THIS WEEK

The paper trail

Scientists must embrace funding-agency efforts to track research outputs and encourage open access to the literature.

Last week, the world's research-funding agencies signalled a welcome desire for wider access to published papers. The Global Research Council, a voluntary discussion forum that takes input from hundreds of funding agencies in regional meetings around the globe, released an action plan for promoting open access — although specific policies are left up to individual agencies (go.nature.com/gonk6w).

Scientists need this top-down push. Individually, they have proved reluctant to make their papers freely available, despite the determined efforts of open-access campaigners. For example, although the Wellcome Trust in London, one of the world's biggest biomedical research charities, has since 2005 provided an open-access mandate and the money to support it, by last June only 55% of the research papers that it funded were open access. Most of those had been uploaded into repositories by publishers, rather than by researchers.

The Global Research Council discussions and action plan have made it clear that when push comes to shove, most agencies lack the will to fund 'gold' open access — in which the author pays for a paper to be free to access as soon as it is published. The UK funding councils, which all provide money towards gold open access, are notable exceptions. Others, such as agencies in the United States, have little money to spare, and are agreeing to wait for publishers or authors to make the results of research freely available six months or a year after publication — 'green' open access. Germany has a dual system, in which researchers can apply for funds to make papers immediately available, but universities can also apply for pots of money expressly to support open access. In Brazil, a system whereby the government negotiates with publishers for open access on a national level is being considered. The permutations are endless.

There is no consensus on whether funding agencies should merely encourage researchers to make their work open (through either green or gold routes), or should actively monitor progress and provide sanctions — such as refusing future grants — for non-compliance. To some extent, reluctance to enforce mandates comes from a desire to ease scientists into open access slowly, but it can also signal endorsement of the idea that agencies should just give researchers opportunities and support for open access, ultimately leaving scientists the freedom to do what they want with their papers.

There is one thing on which funding agencies agree, however. To monitor whether open-access mandates are effective, and to share information on those that are, agencies need to track the outputs of their funding better. At the moment, only a few funders, mainly medical ones — such as the UK Medical Research Council, the Wellcome Trust and the US National Institutes of Health — can give a figure for the proportion of papers resulting from their funding that are open access.

So publishers and funding agencies alike are jumping at new ways to track the sources of funding for published scholarly research. In the same week as the Global Research Council released its action plan, the non-profit publisher alliance CrossRef launched a service called FundRef (www.crossref.org/fundref). The initiative provides a standardized way to report funding sources for published research, by adding them to the metadata on online research papers.

In the United Kingdom, funding agencies expect researchers to provide extensive details on the results of their funding through tracking services known as research outcome systems. Small research charities, which do not have the resources to set up infrastructures to track outcomes, are joining the Medical Research Council in a system called Researchfish. Internationally, many groups are examining standard ways of capturing information about open access; others are looking at how to connect outputs and compare different repositories through mechanisms such as re3data.org, a registry of repositories. At the same time, the ORCID system provides unique identifying numbers to track individual researchers' work, and services such as Figshare are helping to make other types of output, such as data, recognizable and citable.

"To monitor whether open-access mandates are effective, agencies need to track the outputs of their funding."

All this means that as funding agencies push for open access, researchers will need to have their outputs tracked as never before. They should embrace this as a chance both to show off their publications and to acknowledge the support networks that fund their work. ■

Moral authority

Research must be seen to be accountable, even if that means hanging on to redundant reviews.

All scientists must contend with regulation and bureaucracy, despite their frequent complaints that such processes stifle and slow their work. US researchers in gene therapy perhaps feel the pressure of red tape more than most. Is now the right time to ease that burden?

The US Institute of Medicine wants to find out. This week, it kicked off a review of an oversight committee that many in gene therapy argue is redundant. They might be right, but when it comes to medical ethics, it is not enough for scientists to do the right thing — they must also be seen to do so.

The Recombinant DNA Advisory Committee (RAC) was set up within the National Institutes of Health (NIH) in 1974 as a direct response to public concerns about the ethics and safety of research involving lab-assembled DNA sequences. After devising guidelines

for this research, the committee gained the power to approve or reject proposed experiments in humans. As the field has grown, so has the number of experiments that must be given the green light by the RAC, including those that fall under the umbrella of gene therapy.

The problem, as gene therapists see it, is that many other parallel regulatory hurdles have been erected in the meantime. The US Food and Drug Administration (FDA) must approve clinical trials of gene therapy in humans. And institutions have their own biosafety committees and institutional review boards.

Naturally, the RAC disagrees from time to time with the findings of the other scrutineers. When it does, the delays can drag on.

Enough, said the American Society of Gene and Cell Therapy last March. It told the NIH that in recent decades, more than 1,000 gene-therapy clinical trials have been conducted. The worst fears of the public — that gene therapy would lead to alterations of the human genome, or to the release of genetically modified super microbes — have not come to pass. The society told the NIH that the RAC should no longer review individual gene-therapy protocols, and should instead "identify new areas of research that require a public forum for discussion and review".

The gene-therapy field is not free from the risk of adverse events, but the RAC has never claimed to be able to prevent them. Instead, it has had a crucial role in helping the field to learn from setbacks.

After the death of US teenager Jesse Gelsinger in a gene-therapy trial in 1999, for instance, the RAC adopted rules that compel investigators to report all serious adverse events that occur during gene-therapy trials. When children who had been cured of severe combined immunodeficiency by gene therapy developed leukaemia in 2001, the RAC investigated how gene therapy might have contributed to the problem and recommended action to stop it recurring.

Researchers feel strongly that RAC review of individual protocols no longer serves an important purpose, and they resent being at the mercy of the committee's ability to call them out on questions that they often perceive as tangential to their research. Yet it is a tricky time to argue that any public-review process in medical research should be scaled back. Witness the storm of public outrage in the past decade over pharmaceutical companies' failure to report side effects of drugs for conditions from diabetes to depression (see *Nature* **431,** 122–124; 2004). Long after these drugs were approved, it was revealed that their sponsors had held back crucial information about their safety.

"It is a tricky time to argue that any public-review process in medical research should be scaled back."

Gene-therapy clinical trials have proceeded with an unusually high degree of openness, and that has been crucial in helping the field to gain public confidence and acceptance. Such a role should be emulated in other areas of research, rather than eliminated. Apart from the RAC review, none of the oversight required for gene therapy is public. FDA review includes public meetings, but the formal review process allows investigators to keep many of their data secret.

The question is how to preserve the openness that the RAC has enabled without bogging down progress. Perhaps now is the right time to scale back the RAC's purview, but it will be imperative to do so while maintaining the committee's moral authority. That position may feel burdensome, but without it, the field could not have got to where it is today. ■

ANNOUNCEMENT

Nature papers enhanced

As the requirements for data presentation in research papers have grown, *Nature*'s space limitations have remained tight, so more and more essential displayed information has been relegated inappropriately to our Supplementary Information sections. Hard on the heels of our relaxation of constraints on our online Methods sections (see *Nature* **496,** 398; 2013), we are now significantly increasing the number of figures integral to the paper in its online and PDF versions. From July, *Nature* will introduce a new component to its research papers. This new category, called Extended Data (see go.nature.com/tp4vu3), will provide the online reader with immediate access to many display items (figures and tables) previously buried in the Supplementary Information PDF. From now on, most papers submitted to *Nature* can take advantage of this enhancement.

Extended Data display items will be referred to in the print version of the paper, but will be available only online (as is also the case with our full Methods sections). Individual Extended Data display items will be easily accessible by clicking on a call-out in the HTML version of the paper, generating a pop-up box containing the display item and its accompanying legend. Furthermore, the Extended Data display items will be appended to the end of the online PDF, so that the print paper, full Methods section and Extended Data section will be available in one document (see go.nature.com/gb5p6r for a breakdown of the composition of a *Nature* research paper).

Extended Data will not normally contain more than ten individual display items (figures and tables) in addition to the limits set for the printed version of the paper (typically four and five display items for Letters and Articles, respectively). Authors are encouraged to combine appropriate Extended Data figures into multi-panelled figures in order to meet this limit. Each display item should fit onto one page, ideally with its legend or footnote directly below.

The Extended Data display items will be peer reviewed but, like current Supplementary Information, will not be edited in-house. At final submission the Extended Data display items should be generated at the same quality as the figures for the print paper, although there will be differences in formatting (see go.nature.com/zmitgz for a full formatting guide).

Extended Data display items can be used to present essential information relating to the Methods section.

The Supplementary Information section will remain as part of the online-only content, comprising material directly relevant to the conclusion of a paper that cannot be included in the printed version for reasons of space or medium (for example, video clips or sound files). However, this section should no longer contain figures or tables unless there is an exceptional justification (for example, if information is best presented in an Excel file).

From the beginning of July, editors will ask authors who have been invited to revise their papers after the first round of peer review to reformat their papers for consistency with Extended Data. In addition, editors will identify papers at later stages in the editorial process (up to and including the final revision) that might be easily reformatted to include Extended Data display items, and invite authors to revise their papers accordingly. Eventually, all new submissions to *Nature* will be required to comply with this formatting of research papers. The result will be a higher standard of data presentation within the online-only versions of the paper, which will be to the benefit of our readers. ■


GRETCHEN MILLER

US clinical-research system in need of review

An imminent rethink is required on the country's approach to government-supported health and pharmaceutical studies, says **Arthur J. Ammann**.

It is time for a far-reaching and comprehensive review of the way US government-backed clinical research is funded and approved. Ethical reviews of much of this work are currently inadequate and problems come to light too late.

In the most recent example, the US Department of Health and Human Services (HHS) judged that researchers carrying out a study on optimal oxygen administration in infants of low birth weight had failed to fully inform parents about the risks involved.

Clinical researchers defend such studies. But it is clear that too many institutional review boards (IRBs), which give clinical studies the green light, do not have the expertise to thoroughly review the science and ethics of complex clinical trials.

Ethical abuse was certainly more common in the past, and modern science likes to think that it has cleaned up its act. There was an outcry after it was revealed that the US Public Health Service deliberately exposed mentally incapacitated patients, prisoners, sex workers and soldiers to syphilis and gonorrhoea in Guatemala in the 1940s. Officials, including Francis Collins, director of the US National Institutes of Health (NIH) in Bethesda, Maryland, insisted that such unethical studies would be impossible today. Yet a report from the Presidential Commission for the Study of Bioethical Issues in Washington DC challenged that denial. The commission concluded that it "cannot say that all federally funded research provides optimal protections against avoidable harms and unethical treatment".

THE PROMISED IMPERVIOUS WALL OF ETHICAL **PROTECTION** IS RIDDLED WITH **CRACKS.**

On closer examination, the promised impervious wall of ethical protection is riddled with cracks. Progress in clinical research, complicated by the reach of science across borders, has outpaced the ability of researchers and IRBs to make fully informed decisions. The problems are detailed below.

There is inadequate expertise: the composition of IRBs has not kept pace with the complexities of ethics and science. Expert opinions are often derived from individuals who lack sufficient expertise to make an informed decision.

There can be conflicts of interest: individual IRB members may gain salary, health and retirement benefits from approval of research studies conducted at their institutions, which may also make gains.

Exclusivity issues: the design and ethical review of federally funded research is often undertaken by a homogeneous group of individuals with congruent interests at the same or similar academic institutions. Individuals from the public, advocacy groups and non-academic organizations are often excluded. When people from these groups publicly voice their concerns, their views are vilified in academic publications as impeding future advances in research, or even ignored.

➲ **NATURE.COM**
Discuss this article online at:
go.nature.com/sdclni

Marked increases in funding: the NIH budget for research in 2011 was more than US$30 billion. Large amounts of money can distort priorities for research and shift the focus away from urgent public-health needs on the basis of the belief that all research products merit clinical evaluation. The number of products in the therapeutic pipeline is rising and there is no informed method for prioritizing those which should move into clinical research. This increases the risk for people who participate in research.

Increased cost of clinical research and fewer treatment-naive individuals (those who have not been treated with any drugs of the class in question) in the United States: the number of research participants required to obtain statistically significant results for new products has increased drastically because of the need to compare these products with ones that are known to work. A 'mining' approach to obtaining treatment-naive people for research in poor countries has evolved, enlisting vulnerable populations. In some circumstances, the stated benefit for the individual may be limited to the future good of 'mankind', a concept not easily understood in cultures in which health care is deficient. The shift to resource-poor countries is often accomplished by reducing standard of care, exaggerating potential benefits, the use of inferior treatment comparisons and the enrolment of vulnerable people not fully informed of their legal or ethical rights. Although the use of such practices has previously seen pharmaceutical companies criticized, they are increasingly used in academic circles to justify clinical trials funded by the federal government.

A common defence is that breaches of ethical and scientific guidelines are rare. But the Presidential Commission's conclusion clearly states that there is a problem. And the World Medical Association has recently called for revisions of its ethical guidelines, emphasizing that concerns are widespread.

These issues must be resolved before the cracks become fissures. The HHS, the NIH and universities must acknowledge that the current research-approval process is flawed and requires an urgent, comprehensive review that should include experts and leaders from outside the academic community. This review must assess and make recommendations on how research priorities can be established, the means to select the most deserving products for clinical trials, how to expand IRB membership to include greater scientific and ethical expertise, how to minimize conflicts of interest and how to increase public input into decision-making for clinical research. ■

Arthur J. Ammann *is founder of Global Strategies and clinical professor of paediatrics at the University of California, San Francisco Medical Center, San Francisco, California, USA.*
e-mail: arthur.ammann@globalstrategies.org

RESEARCH HIGHLIGHTS

Selections from the scientific literature

CONSERVATION BIOLOGY

The bees are back in Europe

Declines in European pollinator diversity during the twentieth century seem to have slowed following the implementation of environmental policies.

Luísa Gigante Carvalheiro at the University of Leeds, UK, and her colleagues examined historical surveys of native plants and pollinators — bees, hoverflies and butterflies — in the United Kingdom, the Netherlands and Belgium. In general, the number of different native species at a given location, or species richness, declined before 1990. However, after 1990, the decline slowed for most taxa and regions; in the Netherlands and the United Kingdom, local bee richness actually increased.

Although other factors such as climate probably have some role, policies that came into effect after 1990 could be benefiting European species, the authors say.

Ecol. Lett. **http://dx.doi.org/10.1111/ele.12121 (2013)**

IMAGEBROKER/FLPA

FISHERIES

Divers soar after net ban

Breeding populations of diving birds rose after certain fisheries that relied on curtain-like gillnets closed in 1992.

Paul Regular at the Memorial University of Newfoundland in St John's, Canada, and his team analysed seabird census data collected before and after the collapse of fish populations prompted the closure of cod and salmon fisheries in eastern Canada. The team compared the numbers of diving birds such as auks and gannets, which can get entangled in gillnets, with those of surface feeders such as gulls, which thrive on fishery waste. The number of diving birds increased after the nets were removed, whereas scavenger numbers decreased. In particular, growth in the common murre (*Uria aalge*, pictured) population was much higher during the 2000s than the 1970s, whereas the herring gull (*Larus argentatus*) population, which grew in the 1970s, shrank from 2000 to 2010.

The work provides much-needed data to support the theory that fisheries by-catch seriously affects populations of large animals.

Biol. Lett. 9, **20130088 (2013)**

For a longer story on this research, see go.nature.com/9lqmsd

GENOMICS

Spruce shotgun sequencing

Multiple technologies have allowed researchers to piece together the highly repetitive genome of the white spruce (*Picea glauca*, **pictured**), one of the biggest assembled.

Like that of the Norway spruce (*Picea abies*), the draft genome, at more than 20 billion base pairs, is many times larger than the human genome. Inanc Birol at the Genome Sciences Centre in Vancouver, Canada, and his colleagues modified commercial platforms to read longer DNA fragments.

The researchers also used software that relies on parallel computation of overlaps between fragments to determine larger stretches of sequences.

The combined strategies allowed the team to stitch together a genome out of pieces averaging more than 20,000 base pairs. Assembling shotgun sequencing data from scratch can be cost-effective even for gigantic genomes, the authors say.

Bioinformatics **http://dx.doi.org/10.1093/bioinformatics/btt178 (2013)**

WAYNE LYNCH/ALL CANADA PHOTOS/CORBIS

PHYSICAL CHEMISTRY

Square-packed beads

Small beads scattered in liquids can pack into square arrays, an unusual arrangement for floating objects.

Particles at the boundary of a liquid — such as bubbles at the surface of a soft drink — usually clump together in space-saving hexagons to minimize disruption to the surface tension of a liquid. Jasper van der Gucht and his group at Wageningen University in the Netherlands created an experimental set up to explore what would happen if the boundary was curved. They deposited oil droplets on a glass slide, added a layer of water and placed micrometre-sized plastic beads on top. The spheres clustered at the

interface between the oil and water. The team could control the curvature of the interface by changing how oil droplets were attached to the glass, and could generate forces to make the beads group in squares.
Proc. Natl Acad. Sci. **http://dx.doi.org/10.1073/pnas.1222196110 (2013)**

NEUROSCIENCE

Romancing the histones

Pair-bonding in monogamous prairie voles (*Microtus ochrogaster*, **pictured**) is linked to chemical modifications of DNA-packaging proteins in the animals' brains.

Zuoxin Wang, Mohamed Kabbaj and their team at Florida State University in Tallahassee studied the brain chemistry of females as they interacted with males. The researchers focused on enzymes that coordinate epigenetic marks on histones, the protein complexes that coil up DNA and regulate gene expression. Females injected with an inhibitor of these enzymes developed a stronger preference for a random male that had been previously placed in their cage than females not injected with the inhibitor. The inhibitor boosted production of receptors for two hormones linked to sexual and maternal behaviours; similar changes are caused by mating.

The authors say these findings are the first to pair the chemistry of coupling with histone regulation.
Nature Neurosci. **http://dx.doi.org/10.1038/nn.3420 (2013)**
For a longer story on this research, see go.nature.com/fwah2e

ZUOXIN WANG

PARTICLE PHYSICS

A meson unmirrored

A fourth type of subatomic particle shows imperfect symmetry with its antiparticle.

Charge conjugation parity symmetry holds that an experiment should be indistinguishable from its reflection in a mirror if all particles are replaced with their antiparticles. Since the first exception to this was found in 1964, physicists have been hunting for more particles that violate the principle.

A major collaborative effort co-led by Vincenzo Vagnoni at the Large Hadron Collider at CERN, Europe's particle-physics laboratory near Geneva in Switzerland, identified an asymmetry by tracking decays of the subatomic particles B^0_s mesons, which are made up of elementary particles called strange quarks and antibottom quarks. The identification of 'CP violation' in these mesons confirmed a prediction of the standard model of particle physics. However, more-drastic violations are required to explain why matter dominates over antimatter in the universe.
Phys. Rev. Lett. 110, **221601 (2013)**

CLIMATE SCIENCE

More rain in ozone's absence

Ozone loss in the stratosphere over Antarctica has increased rainfall in subtropical parts of South America.

Summers in Uruguay, Paraguay, southern Brazil and northern Argentina became markedly wetter over the second half of the twentieth century. To pin down whether this was due to the impact of ozone on atmospheric circulation and precipitation in the region, Paula Gonzales and her group at Columbia University in New York contrasted simulations with and without ozone depletion using six climate models.

Although results were mixed regarding the impact of greenhouse gases, simulations consistently reproduced observed rainfall trends more accurately when ozone loss was included.

Precipitation in the region should stabilize or decrease as ozone over the Antarctic recovers, the authors suggest.
Clim. Dyn. **http://dx.doi.org/10.1007/s00382-013-1777-x (2013)**

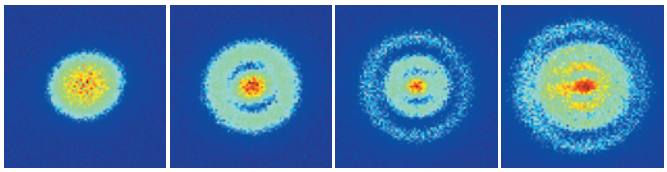
QUANTUM PHYSICS

Direct view of atomic orbitals

Electron orbitals of excited hydrogen atoms can be observed directly.

Orbitals lie outside the nucleus and their properties are described by mathematical wavefunctions. These functions are difficult to study because measuring observable components can destroy other quantum features. Aneta Stodolna at the FOM Institute for Atomic and Molecular Physics in Amsterdam, Marc Vrakking at the Max Born Institute in Berlin and their colleagues designed a quantum microscope to study hydrogen orbitals directly. Their system used tunable lasers to excite electrons in a hydrogen atom placed in an electric field. An electrostatic lens then stretched and magnified the orbitals — without disturbing the internal structure — until individual electrons hit a detector. After recording about 50,000 electrons, the team produced images to show the structure of the electron orbital (**pictured**) of atoms at different excited states.
Phys. Rev. Lett. 110, **213001 (2013)**

ANETA STODOLNA/MARC VRAKKING



COMMUNITY CHOICE

The most viewed papers in science

DEVELOPMENTAL BIOLOGY

Thymus development conserved

HIGHLY READ on dev.biologists.org in April

The genetic mechanisms that regulate how an organ of the immune system forms are conserved between mice and humans.

Tucked inside the ribcage and near the heart, the thymus screens out self-attacking white blood cells. Although molecular and gene-expression patterns that govern the formation of the organ can be readily studied in mice, human studies are generally limited to examining tissue samples under a microscope. Thus, similarities between the species are difficult to assess. Clare Blackburn at the University of Edinburgh, UK, and her team used genetic analysis on tissue from human fetuses. They showed that, in contrast to previous findings, the thymus derives from the same embryological structure in humans as it does in mice. In addition, a collection of genes critical to the organ's development are expressed at similar times in both species, further validating the mouse as a model for thymus development
Development 140, **2015–2026 (2013)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Wandering wheat

Unapproved transgenic wheat has been found on a farm in Oregon, the US Department of Agriculture announced on 29 May. The strain, which bears a transgene that confers resistance to the herbicide glyphosate, was field-tested in 16 US states between 1998 and 2005 by Monsanto, an agricultural-technology company based in St Louis, Missouri (see *Nature* **497,** 24–26; 2013). Regulators are investigating how the wheat escaped but said that it does not pose a risk to food safety. No transgenic wheat varieties have been approved for sale or commercial production in the United States.

Cancer targets

The US Food and Drug Administration announced the approval on 29 May of two drugs for use against advanced melanomas harbouring mutated forms of a protein called BRAF that fuels tumour growth. One drug, called Mekinist (trametinib), is the first cancer drug to target MEK, a protein which is activated by mutated BRAF. The other drug, called Tafinlar (dabrafenib), targets BRAF directly. Both drugs were developed by the London-based firm GlaxoSmithKline. See go.nature.com/ltmmsk for more.

Fish deal decided

European politicians agreed last week on an updated policy governing European Union fisheries. The deal will attempt to restore fish populations to healthy levels and reduce the discarding of unwanted catches. Researchers and conservationists had complained for years that previous rules allowed overfishing. See page 17 for more.

F. PALADIN/FAO

Taking stock of rinderpest

The World Organisation for Animal Health (OIE) in Paris decided last week that it will require its 178 member countries to report annually on any stocks of the rinderpest virus. In 2011, the devastating cattle disease became the first pathogen after smallpox to be globally eradicated (pictured is a person checking for mouth lesions caused by the virus). Informal surveys suggest that at least 40 labs in some 20 countries still hold rinderpest, creating the risk of an accidental or deliberate reintroduction of the pathogen (see *Nature* **488,** 15; 2012). Keith Hamilton, an OIE official, says that the new rule should provide "a more accurate picture of how much virus is still being held in labs", and will aid the containment and destruction of rinderpest.

FACILITIES

SESAME seeded

A funding boost was received last week for SESAME, a synchrotron facility in Jordan intended to promote peace and scientific collaboration between Middle Eastern countries. Italy pledged €1 million (US$1.3 million) in its proposed budget and the European Commission will chip in €5 million for magnets from CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Despite contributions already promised by Jordan, Iran, Israel and Turkey, a further $8 million will be needed to start up the facility in 2015 as planned (see also go.nature.com/5pldwq).

Marine lab ties

On 1 June, members of the Marine Biological Laboratory in Woods Hole, Massachusetts, voted 158 to 2 in favour of an alliance with the University of Chicago in Illinois. The tie would help the 125-year-old lab to dodge further financial hardship after facing a shortfall of nearly US$6 million in its 2012 operating budget. The alliance must still be finalized by each institution's board of trustees. See go.nature.com/lukscr for more.

Open drug libraries

Japan's first public–private partnership aimed at tackling infectious diseases in developing countries announced its inaugural raft of research collaborations on 31 May. The Tokyo-based Global Health Innovative Technology Fund was launched in April with a 5-year commitment of more than US$100 million from the Japanese government, several Japanese pharmaceutical companies and the Bill & Melinda Gates Foundation, headquartered in Seattle, Washington. Its first 13 partnerships will allow three international non-profit organizations to search for candidate drugs for malaria, tuberculosis and a host of neglected diseases by accessing the chemical libraries of Japan's leading pharmaceutical firms.

EVENTS

Coronavirus spread

Last weekend, Italy reported its first cases of the novel Middle East coronavirus: a 45-year-old man from Tuscany who had travelled to Jordan, and two close contacts. The

same pattern of (probably limited) human-to-human transmission has been seen with cases exported from the Middle East to Tunisia, France and Britain. Since last September, the World Health Organization has been told of 53 cases, mostly in Saudi Arabia, including 30 deaths. But the numbers of exported infections suggest that some Middle Eastern cases are going undetected.

PEOPLE

Medical prize

Immunobiologist Ruslan Medzhitov has received the first award of a medical research prize from the Else Kröner Fresenius Foundation in Bad Homburg, Germany. The €4-million (US$5.2-million) award — including €500,000 for personal use — will be given every four years. On 5 June, the foundation chose Medzhitov, of Yale University in New Haven, Connecticut, for his work on the links between the innate immune system, which provides fast, non-specific defence against infections, and the adaptive immune system, which provides specialized responses.

Russian leadership

Russia's largest research organization, the Russian Academy of Sciences, elected a new president on 29 May

SERGEI FADEICHEV/ITAR-TASS

for the first time since 1991. Vladimir Fortov (**pictured**), a 67-year-old plasma physicist and former science minister in the Russian government, succeeds mathematician Yuri Osipov as the academy's leader. The agency employs around 45,000 scientists at more than 400 research institutes across Russia. See go.nature.com/9ntyw7 for more.

RESEARCH

H7N9 virus returns

China reported on 29 May its first new case of the H7N9 avian influenza virus in three weeks — a six-year-old boy who fell ill in Beijing on 21 May. Since March, 132 cases have been confirmed in China, including 37 deaths. But the lull in new cases — possibly attributable to closures of live bird markets — might be only temporary. Animal reservoirs and transmission routes for the virus have yet to be unravelled. Researchers last week reported the emergence of drug resistance in H7N9 to oseltamivir, the mainstay treatment for H7N9 flu (Y. Hu *et al. The Lancet* http://doi.org/mqt; 2013).

Tag-along moon

A 2.7-kilometre-wide asteroid that zipped past Earth on 31 May has been found to have its own moon. Radar imaging from the Deep Space Network antenna complex in Goldstone, California, revealed a 600-metre-wide satellite orbiting asteroid 1998 QE2, which flew 5.8 million kilometres from Earth at its closest approach — about 15 times the Earth–Moon distance. Astronomers in Goldstone and at Arecibo Observatory in Puerto Rico will continue to track the asteroid system to better assess the masses and densities of the rocks.

Restricted access

Reverberations from ongoing legal challenges have led the European Medicines Agency (EMA) to turn down scores of requests for clinical-trial data. The agency is attempting to broaden public access to the information it receives from companies seeking drug approval. However, pharmaceutical firms AbbVie of North Chicago, Illinois, and InterMune of Brisbane, California, took the EMA to court earlier this year to prevent the release of some of their data. See go.nature.com/efqspj for more.

Martian minerals

The Mars Express spacecraft has now mapped the distributions of key minerals over almost all of the red planet's surface, the European Space Agency (ESA) announced on 3 June, a decade after the mission's launch. Clusters of hydrated minerals, detected by the spacecraft's OMEGA instrument, reinforce the view that water was present on Mars in its early history. The data suggest potential landing sites for future missions searching for signs of life on the planet, such as ESA's ExoMars mission.

COMING UP

12–15 JUNE
Researchers gather in Boston, Massachusetts, to discuss the cutting edge of stem-cell science at the annual meeting of the International Society for Stem Cell Research.
go.nature.com/cc1mh1

CORRECTION

The Seven Days item 'Lawsuit settlement' (*Nature* **496,** 402; 2013) wrongly stated that Philippe Bois had successfully appealed against a finding of research misconduct. In fact, Bois had successfully challenged a judge's denial of his request for a defence hearing. In his court settlement, Bois denied that he committed research misconduct, but agreed not to appeal further the findings that he had done so.

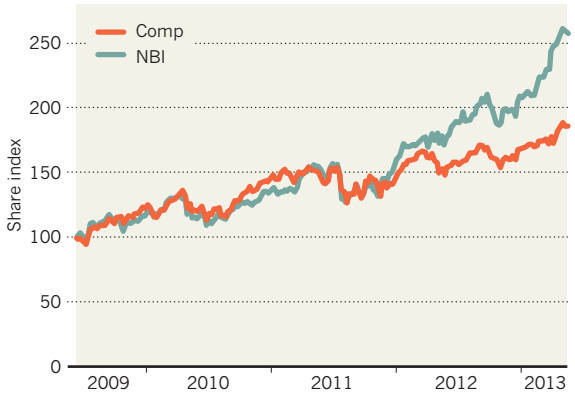
NATURE.COM
For daily news updates see:
www.nature.com/news

TREND WATCH

Two biotechnology firms went public on US markets in the past week, bringing the year's total to 17 and boosting hopes that public markets are again welcoming the industry. Epizyme in Cambridge, Massachusetts, which focuses on cancer treatments, and Kamada, a pharmaceutical firm in Ness Ziona, Israel, entered the market during a boom. Over the past year, the NASDAQ Biotechnology Index, an amalgam of biotech and pharma stocks, has far outpaced the NASDAQ Composite (see chart).

BIOTECH BOOMING

The NASDAQ Biotechnology Index (NBI) is rising much faster than the NASDAQ Composite (Comp).



SOURCE: NASDAQ

NEWS IN FOCUS

MARK LENNIHAN/AP/PRESS ASSOCIATION IMAGES

New York's hurricane-battered Steeplechase Pier will be rebuilt in plastic and concrete.

POLICY

'Plastic wood' is no green guarantee

Researchers question benefits of tropical-wood substitute.

BY JEFF TOLLEFSON

Ishmael Tirado watches as his fellow construction workers rebuild the Steeplechase Pier, a central feature of New York's iconic Coney Island boardwalk. Planks of tropical ipê wood that were torn asunder by last year's Hurricane Sandy lie in grey stacks behind him, ready to be scrapped or recycled, but fresh boards are tellingly absent. When the pier reopens this summer, visitors will encounter a shiny expanse of recycled plastic jutting out to sea on a platform of steel-reinforced concrete. "I think it's a good idea," Tirado says. "It's more durable, and we are saving trees."

Michael Bloomberg, New York's mayor, would probably agree. He promised in 2008 to reduce the city's dependence on tropical hardwoods such as ipê (pronounced 'ee-pay'), and the city has since shifted towards concrete and plastic building materials. Many municipalities and consumers are making a similar choice as they build and maintain outdoor structures. But some researchers fear that a knee-jerk shift away from tropical timber could backfire on the environment.

"If it's sustainable, the timber trade is generally a good thing," says Duncan Brack, an environmental-policy analyst at Chatham House, a think tank in London. "There's a real danger of pushing people towards things with higher environmental impacts."

The scant data available suggest that 'plastic wood' — typically a composite of waste wood and plastic — exacts a higher climate-change cost than natural wood, which has the benefit of pulling carbon dioxide out of the atmosphere as it grows. One 2011 study, funded by the timber industry but independently peer-reviewed, found that the greenhouse-gas emissions from the manufacture of plastic wood are nearly three times higher than those from the production of chemically treated cedar[1].

Data from the Consortium for Research on Renewable Industrial Materials, a public–private partnership based at the University of Washington in Seattle, suggest that emissions from plastic-wood manufacture are 45–330% higher than those of redwood production, depending on whether the plastic is recycled and the extent to which it is supplemented with woody material. Yet plastic wood — which is often marketed as eco-friendly and low-maintenance — is growing in popularity. In the United States, it accounts for around 10% of the market for decking, according to the Freedonia Group, a business consultancy based in Cleveland, Ohio (see 'Timber trade').

The shift has also been driven by supply. In New York, the decision to use more concrete and plastic came after officials concluded that there was no natural timber comparable to ipê — which is prized for its strength and durability — available in sufficient quantity to meet the city's needs. Although the upfront building costs are higher, the assumption is that plastics and concrete last decades with little or no maintenance.

But the evidence for plastic timber's durability is thin, in part because the industry arose only about two decades ago, says Jim Bowyer, a wood scientist at Dovetail Partners, a non-profit environmental consultancy based in Minneapolis, Minnesota. The first generation of plastic timber had problems with sagging and rot; newer products are better, but their long-term performance is hard to predict. "I know of one lab that has gotten mushrooms to grow on it," Bowyer says. And without solid data on the lifetimes of different types of plastic timber, it is difficult to assess their environmental impacts, he adds.

Information about tropical-timber production is incomplete but suggests that the industry has an outsize environmental impact. According to the International Tropical Timber Organization, tropical countries supply about 10% of the world's industrial wood, much of it from plantations.

But of the 400 million hectares (more than half of the world's total) of tropical forest used for timber production today, less than 8% is sustainably managed. In many countries, illegal logging still accounts for much of the ▶

production, and many roads built legitimately by logging companies become arteries for illegal agricultural development. As such, logging often serves as a precursor to large-scale deforestation.

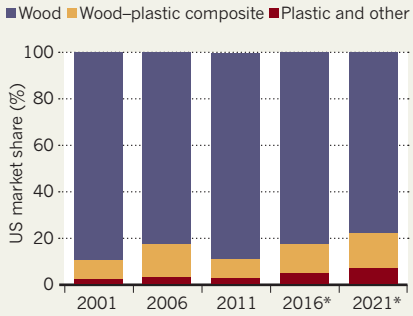
The simple solution is to avoid tropical wood, but that undercuts the market for timber that is sustainably produced, says Doug Boucher, head of the tropical-forestry programme for the Union of Concerned Scientists in Cambridge, Massachusetts. "This kind of quick and relatively unsophisticated response basically hurts tropical countries, at the expense of the developed world," he says.

Industry and governments have taken steps to improve the market for sustainable tropical wood. The Forest Stewardship Council was launched in 1993 to create an independent certification for sustainably produced timber. More recently, many countries have banned the importation of illegally produced timber. The United States instituted such a ban in 2008. Europe this year went a step further by implementing regulations requiring companies that import wood to establish plans to ensure that the imports are legal. The United Kingdom requires all government-procured wood to be certified as sustainable.

The impact of these initiatives is not yet clear, but there are signs of progress. A 2010 study by Chatham House found that illegal logging had dropped by nearly 25% over the preceding decade, as enforcement efforts increased in tropical countries[2]. But corruption remains a global problem. Many illegal timber exports are going to China, where they are blended into the larger industrial supply chain.

TIMBER TRADE

Analysts expect the US appetite for 'plastic wood' decking materials to increase in coming years.



*Market projections.

SOURCE: THE FREEDONIA GROUP

The response to the UK approach suggests a way forward: the law has prompted many major importers in that country to certify the sustainability of their entire supply chains. As much as 80% of the wood entering the United Kingdom now comes with such assurances, says Brack, and the policy may be shaping broader practices. "The UK government is a big enough consumer of things like office furniture, paper and timber for construction that it has quite a large knock-on effect on the rest of the market," he says.

Broad numbers aside, a question facing many governments and consumers is whether ipê in particular is being managed sustainably. Some ipê is certified by organizations such as the Forest Stewardship Council, but the overall production data are uncertain. "Is the use of ipê sustainable? No one really knows," Bowyer says. "We are kind of operating in the dark."

At Coney Island, many residents are fighting to protect the signature wooden boardwalk, and the city has acknowledged the historical significance of natural wood in some locations. On a recent weekend, construction crews were busy replacing small sections of boardwalk near the Steeplechase Pier — with fresh ipê. ■

1. Bolin, C. A. & Smith, S. *J. Clean. Prod.* **19,** 620–629 (2011).
2. Lawson, S. & MacFaul, L. *Illegal logging and related trade: indicators of the global response* (Royal Institute of International Affairs, 2010).

PUBLISHING

Tensions grow as data-mining discussions fall apart

Scientists want to exempt computer-based text crawling from Europe's copyright law.

BY RICHARD VAN NOORDEN

Disagreement between scientists and publishers has grown on a thorny issue: how to make it easier for computer programs to extract facts and data from online research papers. On 22 May, researchers, librarians and others pulled out of European Commission talks on how to encourage the techniques, known as text mining and data mining. The withdrawal has effectively ended the contentious discussions, although a formal abandonment can be decided only after a commission review in July.

Scientists have chafed for years at limitations on computer-aided research. They would like to use computer programs to crawl over thousands or millions of articles and other online research content, extracting data to build up databases or to pick out patterns such as associations between genes and diseases.

But in many parts of the world, including Europe, this sort of use currently requires permission from the content's copyright owner. Even if an institution has paid to access a journal, its academics do not necessarily have permission to mine the text. Publishers, worried that their content might be redistributed for free, tend to block data-mining programs, giving extra licence permissions only on a slow, case-by-case basis (see *Nature* **483,** 134–135; 2012). And although authors can now choose to publish under licences that explicitly allow text mining, that innovation doesn't help text-miners wanting to run programs on decades of pre-existing content.

Rather than struggle through a thicket of different permissions set by publishers, some researchers want Europe to exempt text mining from copyright law — allowing them to run programs on content that they have paid for, and on free content, without fear of copyright breach. Last year, the UK government said that it plans to introduce exemptions for non-commercial purposes. Lenient 'fair use' rights in the United States may already allow text mining, depending on how the law is interpreted.

NATURE.COM
For more on the changing world of publishing, see:
nature.com/scipublishing

"There is an intense debate on this within the scientific and research community, with a large number of scientists pointing at the limits of the current copyright regulatory regime," says Ryan Heath, a spokesman for European Commission vice-president Neelie Kroes. "This is a very serious issue, impacting on scientific excellence and innovation in Europe."

To tackle the issue, last December the commission set up a working group — one of a number under a framework called Licences for Europe — to open discussions about new policies among publishers, researchers, librarians and other interested parties, such as technology companies. In late February, researchers complained in a letter to the commission that

SCIENCE/AAAS

the group was constrained to discuss only text-mining licences, and not changes to copyright law (see *Nature* **495,** 295; 2013) — a restriction that would "make computer-based research in many instances impossible".

"Every researcher I've spoken to thinks licensing is a problem," says Susan Reilly, projects manager at the Association of European Research Libraries in the Hague, the Netherlands. She coordinated the letter that declared the 22 May withdrawal from talks. "There was really no point in us continuing to attend," she says. Other signatories include the non-profit Open Knowledge Foundation in Cambridge, UK, and the National Centre for Text Mining at the University of Manchester, UK.

"Continuing the group under current circumstances doesn't make sense," says Heath. "This is regrettable, but at least the process brought to the fore the major controversies in this area." The European Commission, he adds, "will reflect on the implications and will address the matter at the time of the review of the Licences for Europe process in July".

The European talks had always been conflicted because four different European Union administrative departments were involved — not only the department for research and innovation, but also those for education and culture, for media and information issues, and for Europe's internal market, economy and intellectual-property rights. (The May letter argues that the research department is being squeezed out in favour of the others' interests.)

"Since the Licences for Europe process has not managed to deliver in this area, other ways forward must be explored," says Heath. An analysis under way by the commission's internal-market department on the need for copyright reform may provide impetus for action, should it conclude that changes are needed.

Many publishers say that there are practical, as well as legal, barriers to text mining. Even if the practice were permitted through licences or changes to copyright law, researchers would still need a way to access websites without crippling publisher servers through excess traffic. And publishers want to be able to identify the purpose of the programs crawling their content, especially if mining is for commercial means, so as to decide "what they're willing to allow at what cost", says Sarah Faulder, chief executive of the Publishers Licensing Society in London, an industry body that took part in the talks.

To lower some of these practical barriers, the non-profit publisher collaboration CrossRef hopes to launch technology this year enabling text-mining researchers to agree to terms by clicking a button on a publisher's website.

Discussions may have faltered, but scientists and librarians hope to keep talking to officials, says Reilly. "There's lots of disagreement even among publishers," she says. "Some are open to text and data mining, some are completely frightened of it. They need an informed discussion." ■

A glow-in-the-dark tobacco plant was first engineered by scientists in the 1980s.

SYNTHETIC BIOLOGY

Glowing plants spark debate

Critics irked over planned release of engineered organism.

BY EWEN CALLAWAY

Among the many projects attracting crowd-sourced funding on the Kickstarter website this week are a premium Kobe beef jerky, a keyboard instrument called a wheelharp and a small leafy plant that will be made to glow in the dark using synthetic-biology techniques.

The Glowing Plant project, which ends its fund-raising campaign on 7 June, seeks to engineer the thale cress *Arabidopsis thaliana* to emit weak, green-blue light by endowing it with genetic circuitry from fireflies. If the non-commercial project succeeds, thousands of supporters will receive seeds to plant the hardy weed wherever they wish.

The US government has no problem with this prospect, yet some experts and industry watchers are jittery. They fear that distributing the plants could set a precedent for unsupervised releases of synthetic organisms, and might foster a negative public perception of synthetic biology — an emerging experimental discipline that involves genetically engineering organisms to do useful tasks.

The project, based in the San Francisco Bay Area in California, was conceived as a public demonstration of synthetic biology using gene-writing software and lab-made DNA molecules. The effort also reflects a 'DIY biology' movement that seeks to make biotechnology more accessible to the public. "The central goal of the project is to inspire people and educate people about this technology," says entrepreneur and project co-founder Antony Evans.

He and his colleagues — Omri Amirav-Drory, founder of synthetic-biology software firm Genome Compiler in Berkeley, California, and Kyle Taylor, a former biology graduate student at Stanford University in California — set out to make *Arabidopsis* glow because the feat seemed achievable in a simple garage lab. "There are some people in synthetic-biology circles who would yawn at what we're doing," Evans says.

Making plants glow has been possible since the 1980s, when scientists added a gene encoding the firefly enzyme luciferase to a tobacco plant. When sprayed with the chemical substrate luciferin, the plant glowed temporarily (D. W. Ow *et al. Science* **234,** 856–859; 1986). In 2010, another group engineered a tobacco plant to have its own weak glow, using bacterial genes instead (A. Krichevsky *et al.* ▶

PLoS ONE **5,** e15461; 2010). Also in 2010, a team at the University of Cambridge, UK, created a genetic circuit in bacteria that makes both firefly luciferase and luciferin, so that the bacteria glow continuously (go.nature.com/4nxcao). The Glowing Plant team plans to tweak the genes in that circuit so that they work in plants.

The more than 7,700 project supporters will also be rewarded with stickers, T-shirts depicting glowing plants or light-bulb vases. The effort hit its initial fund-raising goal of US$65,000 several weeks early, and passed the $400,000 mark on 28 May. With the extra cash, Evans and his team will try to create glowing roses too. They are taking no salary, and are borrowing lab and greenhouse space. "It's a really positive signal for synthetic biology that there's this big consensus-level interest in genetically engineered objects," says Mackenzie Cowell, founder of a San Francisco biotech-supply company called Genefoo. He chipped in $250 to the effort.

But Drew Endy, a synthetic biologist at Stanford University, questions how much light the plants will actually be able to emit, given the limitations on a plant's ability to harvest energy from the Sun and convert it back into light. "Never mind the genetic engineering involved — just what does the physics say about the feasibility of the project working out?" he says.

"Is this legal?" asks the project's Kickstarter site, with the reply "Yes it is!" Evans says that he and his team contacted the Animal and Plant Health Inspection Service (APHIS) at the US Department of Agriculture, which regulates genetically modified (GM) plants if plant pathogens are involved in the work. The agency's main concern was whether DNA from the pathogen *Agrobacterium* would be used to insert foreign genes, as GM plant efforts often do. "Regarding synthetic biologics, if they do not pose a plant risk, APHIS does not regulate it," a spokesperson told *Nature*.

To bypass this concern, the Glowing Plant team will use *Agrobacterium* only during preparatory tinkering with the luciferase genetic circuit. When plants are produced for distribution, the team will shuttle the genes into cells using a ballistics-powered device called a gene gun, a process that the agriculture department deems outside its purview (see *Nature* **475,** 274–275; 2011).

Such regulatory runarounds need to be scrutinized, says Todd Kuiken, who studies synthetic-biology issues at the Woodrow Wilson International Center for Scholars, a think tank in Washington DC. Although he has few concerns about streets lined with glowing *Arabidopsis*, he thinks that the lack of oversight of future, riskier projects could prove problematic.

And Allison Snow, an ecologist at Ohio State University in Columbus who studies the risks posed by GM plants, says that it won't do synthetic biologists any public-relations favours if plants make it into the wild. People will be more likely to support synthetic biology, she says, if it is associated with disease treatments or clean biofuels. "This is such a frivolous application," she says (see 'Bioluminescent boom').

Some people are riled already. The ETC Group, a Canadian pressure organization in Ottawa with a history of opposing synthetic-biology applications, launched a "kickstopper" campaign against the project and is looking into legal options to stop it.

Evans says that the team is likely to engineer a type of *Arabidopsis* that survives only if fed a nutritional supplement, reducing the chances of spread. And the team plans to conduct a public dialogue on the project's ethical, legal and environmental issues before shipping any seeds. "This is a fund-raising campaign," he says. "It's not the actual release of the plant."■

GLOWING REPORT

Bioluminescent boom

The Glowing Plant project is not the only foray into publicly available genetically modified organisms. Transgenic zebrafish (*Danio rerio*) that produce a fluorescent protein have been on the market since 2003, although their sale is not permitted in the European Union, Canada, Australia or California. And BioGlow, a commercial venture in St Louis, Missouri, informed the US agriculture department last year of plans to produce light-emitting plants, but the company has made few details public.

GENOMICS

Geneticists push for global data-sharing

International organization aims to promote exchange and linking of DNA sequences and clinical information.

BY ERIKA CHECK HAYDEN

It is a paradox that bedevils genomic medicine: despite near-universal agreement that doctors and geneticists should exchange more data, there has been scant movement towards achieving this goal.

Now, a consortium of 69 institutions in 13 countries hopes to address the problem by creating an organization to enable the free flow of information in genomic medicine. On 5 June, the consortium, which is calling itself the 'global alliance', announced that the organization will develop standards and policies to encourage data-sharing of a person's DNA sequence combined with clinical information. The alliance's founders are basing their model on the World Wide Web Consortium, which in the 1990s established standards for the programming language HTML and spurred the growth of web pages across the Internet.

"This alliance steps into what otherwise might be a real void," says Francis Collins, director of the US National Institutes of Health (NIH) in Bethesda, Maryland, which is a member of the alliance. For example, Collins says, there are no standards for storing genetic sequences or for assessing their accuracy.

NATURE.COM
For more on genetic data-sharing, see:
go.nature.com/5oxmj7

The alliance also hopes to tackle privacy and informed-consent issues that prevent researchers from sharing data, and plans to create a network of cloud-computing platforms and analysis tools in an effort to provide access to the shared data.

A big question for the group is whether it can convince institutions to share their most meaningful data. "The mission is unquestionably worthy," says cardiologist Eric Topol, director of the Scripps Translational Science Institute in La Jolla, California, which has not yet considered joining the alliance. But, he adds, "it means taking the walls down, and that's tricky — because you've got each centre wanting to hold on to its own data, and the loss of control is a very difficult concept".

The effort has gained support from some of the world's most influential sequence-data holders, including the NIH, the Wellcome Trust Sanger Institute in Hinxton, UK, and the BGI (formerly the Beijing Genomics Institute) in Shenzhen, China. David Altshuler, a geneticist at the Broad Institute in Cambridge, Massachusetts, who led an eight-person organizational committee for the project, is keen to add more members. "We're saying, 'This is bigger than any group or institution — let's figure

out how to get it right'," he says.

With the cost of sequencing falling with each passing year, the number of sequenced human genomes is now poised to reach into the millions. But researchers can't gain a complete picture of how genes influence disease unless those data are linked to clinical information and different institutions share data with each other.

Researchers are often reluctant to share this hard-won information, however. And on occasion, because of privacy concerns, they are legally prevented from doing so. That blocks scientists' ability to use the world's collective data to find answers to simple questions, such as how often a particular genetic variant is linked to a disease.

The establishment of technical standards for storage and sharing will go part of the way towards making genomic data easier to share and analyse. But the alliance also hopes to surmount some of the legal barriers by establishing how anonymity is handled and what information needs to be kept secure. Institutions that abide by core principles could then share data even if their policies differed in other, less central ways.

Moreover, the alliance wants to encourage the development of tools to allow patients to maintain control over their own medical and genetic data. Harold Varmus, director of the National Cancer Institute (NCI) in Bethesda, suggests that institutions should be able to tag their data so that it is accessible only for certain studies — a step that is "going to be incredibly important", he says.

Some major genomic-medicine projects have signed up to the alliance, but others have not yet joined, and have limited outsiders' access to their data. That is partly to head off privacy and security concerns, but also because the information is such a valuable commodity (see 'Precious data').

In the future, research funders such as the NIH and NCI could induce more projects to join by asking grantees to abide by policies set by the alliance, Collins and Varmus say. The project's success will depend on the alliance convincing organizations that it is worth giving up some control to gain access to a broader universe of data, says Michael Stratton, director of the Sanger Institute. "We're committed to the idea that sharing data will be central to extracting the maximum amount of knowledge for the benefit of humankind," he says. ■

PRECIOUS DATA

A 'global alliance' of research institutes wants to encourage sharing of linked genetic and clinical data, but not all of the major data holders have joined the project.

Project	Enrolled participants	Joined global alliance?
US Million Veteran Program	213,000	No
Vanderbilt University BioVU	165,000	No
Kaiser Permanente Research Program on Genes, Environment, and Health	430,000	No
UK10K	10,000	Yes
Deciphering Developmental Disorders	12,000	Yes

CONSERVATION

Europe reforms its fisheries

Agreement would set catch limits that are in line with scientific advice.

BY DANIEL CRESSEY

The breakthrough came at around 3 a.m. on 30 May in Brussels, after a marathon negotiating session: the European Union (EU) finally agreed to end overfishing in its troubled waters.

Fisheries scientists say that the deal, which is expected to be approved before the end of the year, could allow fish stocks to recover to their previous bountiful levels, after being driven down by years of overfishing. But short-term restrictions are likely to bring unemployment to some fishermen.

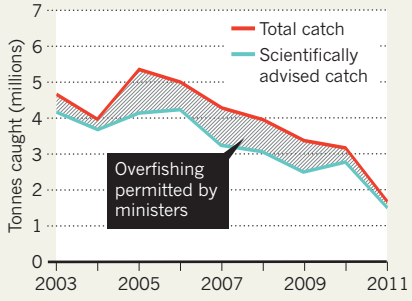
"There is bound to be some short-term pain," says Michel Kaiser, who studies fisheries at Bangor University, UK. "This reform has come about because there was a groundswell of realization that what we had before couldn't go on."

The deal places scientific advice at centre-stage in determining catch limits, as the EU commits to fishing at healthy levels by 2015 "where possible" and by 2020 otherwise. New rules will also be phased in to reduce ecologically damaging 'discards' — the practice of throwing fish caught in the pursuit of other species back into the sea, with the vast majority dying in the process.

For years, scientists have warned that more fish were being caught than was sustainable, owing to a flawed 'Common Fisheries Policy' (CFP), which governs commercial fishing in European waters. Government ministers set higher catch limits for cod, haddock and some other species than scientists considered wise (see 'A waning haul').The latest agreement, which has been several years in the making, is backed by the three arms of European government: the commission, parliament and council.

A WANING HAUL

European ministers have consistently ignored scientific advice in setting catch limits for 107 fish stocks in the northeast Atlantic fishery.



SOURCE: WWF, EU

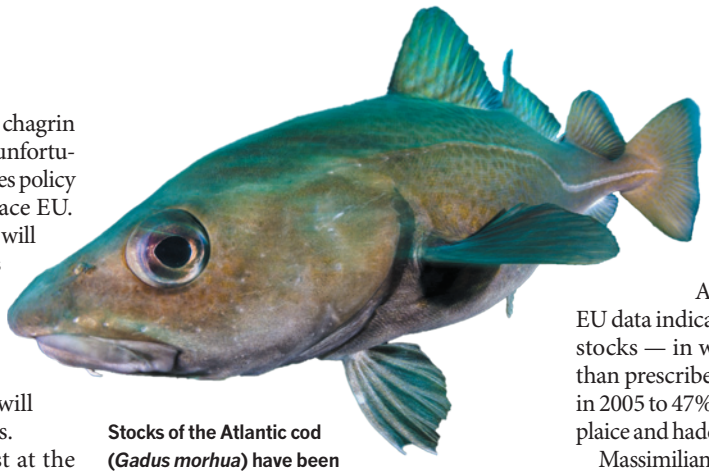
Parliament had been pushing for a thorough reform of the CFP to put catches in line with what science says is sustainable, whereas the council — made up of ministers from EU member states — had been less amenable to radical change.

Environmentalists are generally pleased with the deal's main thrust: a commitment to fishing at maximum sustainable yield (MSY), the largest catch of a particular species that can be taken indefinitely without harming the main population. Scientists have two measures for MSY, obtained using mathematical models created with data from catches by commercial and research vessels: the overall biomass of a species needed to maintain MSY (B_{MSY}) and the annual amount of fish taken from that species that will still allow the species to reach B_{MSY} (F_{MSY}). Fishing at a higher level than F_{MSY} means the fishing is unsustainable in the long term. Environmentalists prefer B_{MSY} to F_{MSY} as a target, because reaching the former would show that a stock has actually recovered, whereas fishing in line with the latter indicates that a stock is on the road to recovery.

The EU agreement would set catch limits at F_{MSY} by 2015 where possible, and by 2020 in other cases. It has also promised to move to ▶

B_{MSY}, but without a firm date, to the chagrin of conservationists. "That's one of the unfortunate things," says Saskia Richartz, fisheries policy director for Brussels-based Greenpeace EU. Richartz also worries that EU ministers will have the final say in setting catch limits and may not stick to the science. "It now says in the text very clearly [ministers] must stick to scientific advice," says Richartz. But "it remains hope rather than certainty" that ministers will honour the F_{MSY} targets set by scientists.

Rainer Froese, a marine ecologist at the GEOMAR Helmholtz Centre for Ocean Research in Kiel, Germany, is also not entirely pleased with the agreement. He says that the council has won a loophole in the 'discard ban', in that some fishermen will still be able to throw back up to 5% of their catches. Critics also say that the 5% exemption will make excessive discarding difficult to enforce, because it will be hard to prove that fishing operations, caught in the act of throwing animals back into the sea, are exceeding their quota.



ALEX MUSTARD/NATUREPL.COM

Stocks of the Atlantic cod (*Gadus morhua*) have been decimated in recent years.

Froese also worries about the willingness of member states to set catch limits in line with F_{MSY}, and says that there will be pressure on scientists to increase their estimates of F_{MSY} in a way that benefits the industry. His own research suggests that the fisheries for some stocks, such as the North Sea cod, will need to be closed altogether for several years before the population can recover.

Other experts are more positive about the reform, and note that catches in recent years have already moved closer to scientists' advice. There are even signs that some northeast Atlantic stocks are bouncing back: EU data indicate that the number of overfished stocks — in which more animals are caught than prescribed by F_{MSY} — dropped from 94% in 2005 to 47% in 2012. Some stocks of herring, plaice and haddock are now fished at F_{MSY} levels.

Massimiliano Cardinale, a fisheries researcher at the Swedish University of Agricultural Sciences in Lysekil, says that although some stocks are recovering, the big challenge will be recovering the over-exploited and commercially important top predators such as cod and tuna. Bringing them back would reshape entire ecosystems off Europe's coasts, he adds.

This will not happen by 2015, and probably not by 2020, says Cardinale, but with a bit more time "the ecosystem might look more like it should do". ■

WORLD HEALTH ORGANIZATION

Agency gets a grip on budget

Reforms increase flexibility and shift spending towards non-communicable disorders.

BY DECLAN BUTLER

Just three years ago, the World Health Organization (WHO) was in deep financial trouble, with a US$300-million deficit. Today the agency's future looks healthier. Last week, the World Health Assembly — the annual gathering in Geneva, Switzerland, of health ministers of the WHO's 194 governing member states — voted in favour of major budgetary reforms that look set to put the agency on a firmer financial footing.

The agency has also taken action to prune and prioritize its work, which critics say has long been spread too thinly. Taken together, the budget and streamlining reforms "are clearly an effort, that is visible and tangible, to get their house in order at multiple levels", says Barry Bloom, a global-health expert at the Harvard School of Public Health in Boston, Massachusetts, and an ardent advocate of WHO reform.

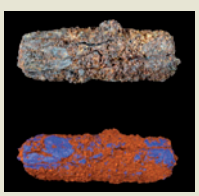
The $3.98-billion budget approved by the assembly for 2014–15 shows zero growth on the WHO's $3.96-billion budget for 2012–13, and marks a slight decrease when inflation is taken into account. The numbers are in line with a worldwide flatlining of spending on global health after a decade of rapid growth that saw much public-health spending shift to new players (see 'Peak health').

This freeze has forced the agency to make some hard choices. The budget breakdown shows a shift away from infectious diseases — with a $72-million cut, taking expenditure down to $841 million — towards work on non-communicable disorders such as cardiovascular disease and cancer. These received a $54-million increase, to $318 million. The changes correct what experts say has long been an inappropriate skew in the organization's budget. They also tie in with UN-wide plans for a global push to reduce the burden of non-communicable diseases, in particular by reinforcing health-care systems in poorer countries where these ills are often neglected. But with no increase in the budget, cuts in some sectors are inevitable if other sectors are to grow.

In a world facing outbreaks of H7N9 influenza in China and a novel coronavirus in the Middle East — both potential pandemic threats — some public-health experts are concerned by a 51% spending cut for

MORE ONLINE

TOP STORY



Ancient Egyptian relics were made with iron from meteorites go.nature.com/cudsuq

MORE NEWS

- Mars robotic mission measures radiation to which humans would be exposed go.nature.com/tweanr
- Gun-violence researcher turns to crowd-funding go.nature.com/trfc4n
- Antibiotic-resistance researcher talks outreach go.nature.com/qrsy66

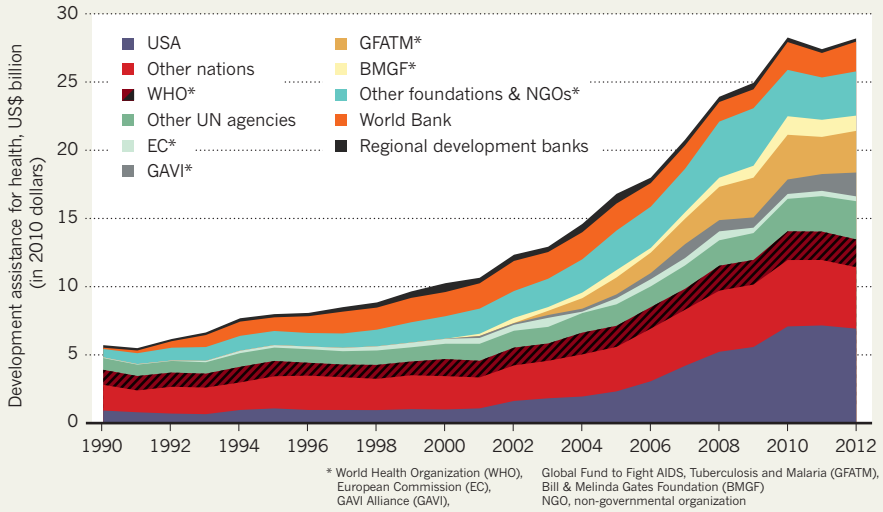
NATURE PODCAST

Atoms with crazy shapes, a time cloak to hide events, and what is really happening to ice sheets nature.com/nature/podcast

OPEN UNIV./UNIV. MANCHESTER

PEAK HEALTH

As contributions by other players grow, the World Health Organization is no longer the dominant force in global health.



* World Health Organization (WHO), European Commission (EC), GAVI Alliance (GAVI), Global Fund to Fight AIDS, Tuberculosis and Malaria (GFATM), Bill & Melinda Gates Foundation (BMGF) NGO, non-governmental organization

SOURCE: *FINANCING GLOBAL HEALTH 2012: THE END OF THE GOLDEN AGE?* (IHME, 2013)

the WHO's 'outbreak and crisis response' — from $469 million to $228 million. Gaudenz Silberschmidt, a senior adviser to WHO director-general Margaret Chan, says that this cut mainly reflects the difficulty of predicting the spending needs for such outbreak-response work. He adds that when crises occur, the WHO will seek emergency funding from member states. "If H7N9 or coronavirus turn nasty, it's obvious that member-state donors will be ready to give more," he says.

The WHO is in fact expanding its work to prepare for, and respond to, outbreaks and other global-health threats, adds Silberschmidt. It is shifting towards helping countries to respond for themselves, rather than depending on the WHO as a global fire brigade. A separate budget line devoted to this — 'preparedness, surveillance and response' — will increase by 32% to $287 million.

The shift stems from a 2007 agreement by the WHO's member states to have a legally binding set of rules on handling outbreaks or other public-health threats of potential global significance: the International Health Regulations (IHR). These rules, which are largely a response to weaknesses seen in some countries' responses to the outbreaks of SARS and H5N1 flu in the early 2000s, oblige countries to put in place a series of measures to enable adequate action when outbreaks occur. The measures include establishing disease-surveillance networks and reporting mechanisms, and installing lab and other core infrastructure.

But a progress report that Chan presented at last week's assembly shows that few countries have met the June 2012 deadline for implementing the measures. "The IHR will never be effective unless that surveillance and lab infrastructure is in place," says Adam Kamradt-Scott, a health-policy researcher at the University of Sydney in Australia.

The biggest change in the WHO budget involves details of a new financial architecture. The agency has long been plagued by the fact that it has total control of only a small part of its budget: monies coming from the membership fees of its 194 states. The bulk — 77% — of the 2014–15 budget comes from voluntary contributions from member states and other donors.

Voluntary donations are usually earmarked for pet priorities. As a result, the WHO's work is pulled in all directions by its donors, often without commensurate funding. Even worse, until now the assembly has approved only the membership-fees component of the budget, whereas the pledged voluntary contributions can vary by as much as 30%, says Silberschmidt, making it difficult to plan.

From now on, the voluntary contributions will be fixed commitments rather than pledges. Another innovation is a rule that allows the WHO to move up to 5% of one budget line to another, providing flexibility in addressing unforeseen needs.

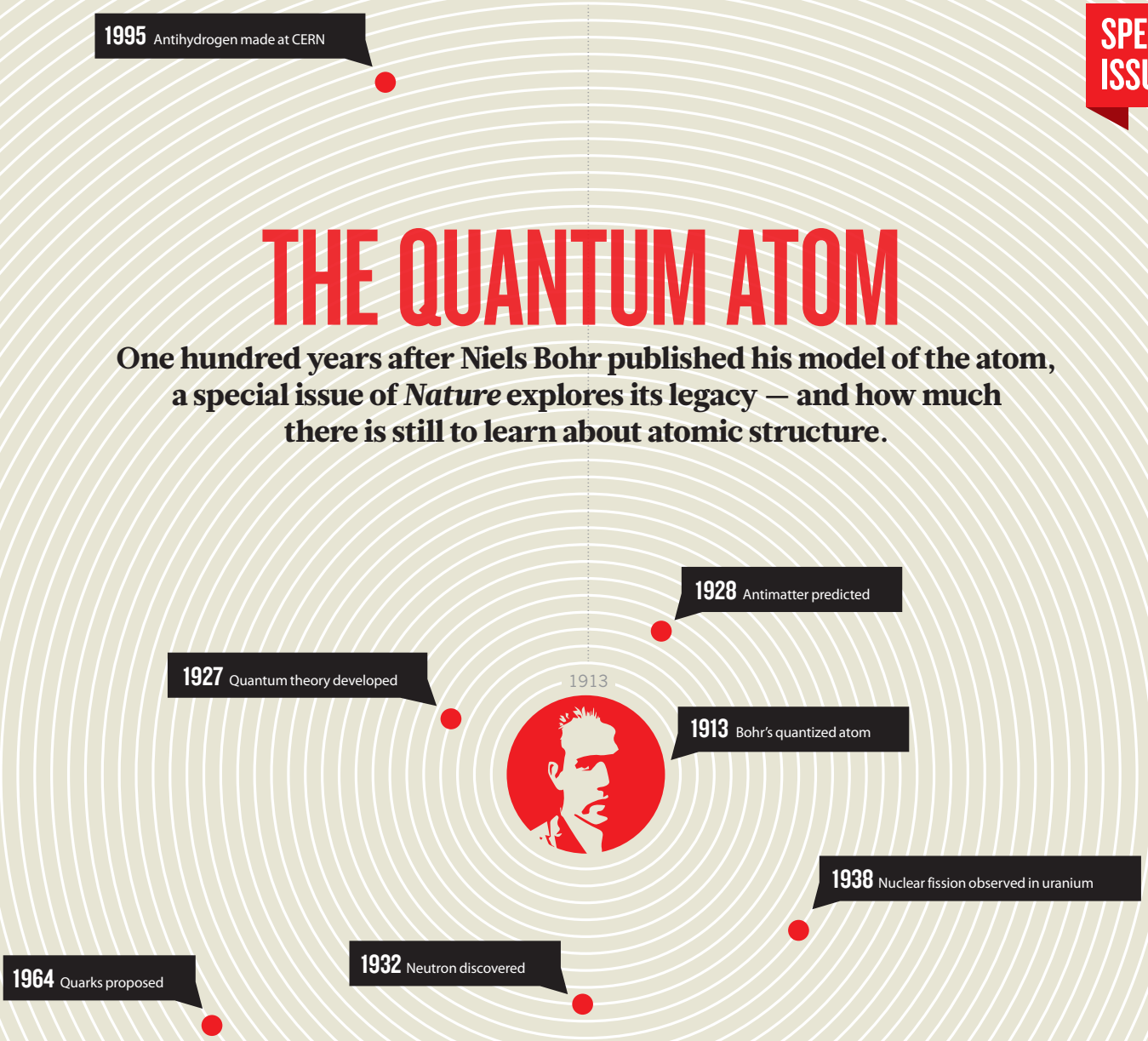
Kamradt-Scott calls the changes "fairly substantial reforms in the WHO's ability to manage its finances". They also make it far clearer to the public just how much money the WHO receives and where that money goes.

But although the budget changes are helpful, they do not necessarily overcome the fundamental problem, says Lawrence Gostin, head of the WHO Collaborating Center on Public Health Law and Human Rights at Georgetown University in Washington DC. Too large a piece of the WHO budget pie comes from voluntary contributions, making the agency's work and policies ultimately reflective of its wealthiest donors, and leaving it scant margin to set its own. "It simply is not sustainable to have wealthy states and foundations control some 80% of WHO's budget," Gostin says. ■

SPECIAL ISSUE

THE QUANTUM ATOM

One hundred years after Niels Bohr published his model of the atom, a special issue of *Nature* explores its legacy — and how much there is still to learn about atomic structure.



July 1913 saw Danish physicist Niels Bohr publish the first of three papers setting out a radical new view of the nuclear atom. His idea — a positively charged nucleus ringed by electrons in orbits of discrete energies — explained the frequencies of light emitted by hydrogen as electrons made leaps between orbits. Quantum rules determined the electrons' energies, preventing the instabilities that had plagued previous mechanical models of atoms.

This special issue of *Nature* explores the origin and legacy of Bohr's quantum atom, a model that has resonated ever since. In 1911, Bohr began a postdoctoral year in England that planted the seeds of his thinking. In a Comment on page 27, historian John Heilbron relates how letters from Bohr to his brother Harald and to his fiancée, Margrethe Nørlund, published this year, chart the dauntless physicist's work with J. J. Thomson and Ernest Rutherford, and his study of the papers of John William Nicholson, which presaged his breakthrough.

The kaleidoscopic nature of the electron is illuminated by physicist Frank Wilczek in a second Comment (page 31). For most practical purposes, electrons behave like simple point particles — but at high energies, they reveal their constituents in showers of quarks, gluons and neutrinos. Physicists are still striving to understand puzzling manifestations of electrons such as coupled states in superconductors and fragments with fractional charges.

Other researchers are testing the limits of the Bohr model by, for example, using powerful X-ray lasers to blast away inner electrons and create 'hollow' atoms. A News Feature explores these and other extreme atoms, including giant, superheavy and antimatter forms (page 22). Such explorations may hit limits on atomic and nuclear size, as two physicists discuss in a News and Views Forum on page 40. Wildly courageous and at ease with ambiguity, even Bohr would have struggled to anticipate the impacts of his vision. ■

THE QUANTUM ATOM
A *Nature* special issue
nature.com/bohr100

ILLUSTRATION: THOMAS POROSTOCKY

EXTREME ATOMS

Physicists are stretching, stripping and contorting atoms to new and bizarre limits.

BY RICHARD VAN NOORDEN

One way to obliterate an atom is to shoot it with the planet's most powerful X-ray gun. Linda Young tried that experiment in October 2009, when she was testing the newly opened X-ray free-electron laser at the SLAC National Accelerator Laboratory in Menlo Park, California. A single pulse from the US$420-million machine packs the same energy as all the solar radiation hitting Earth at that moment, but focused down to one square centimetre. "It will destroy anything you put in front of it," says Young.

When the laser pulse slammed into the neon atoms in that experiment, it made them explode, stripping away each atom's 10 electrons within 100 femtoseconds (1 femtosecond is 10^{-15} seconds). But it was the manner of this destruction that most interested Young, who heads the X-ray science division at Argonne National Laboratory in Illinois. The X-rays first removed the atom's inner electrons, leaving the outer ones in place. For a brief moment, the neon atoms in the path of the laser became hollow.

That exotic form of neon is one of a number of strange species created by physicists intent on contorting atoms. Some teams have inflated atoms to the size of dust particles. Several research collaborations are creating anti-atoms out of antimatter. And others have loaded atomic nuclei with protons and neutrons in the quest to forge new superheavy elements. Some of the experiments aim to investigate atomic structure; others use atoms as the first steps in modelling more complicated systems. They are all descendants of the revolution in atomic theory catalysed by Danish physicist Niels Bohr 100 years ago. But Bohr would have had difficulty imagining how far scientists could go in poking and prodding atoms into such extreme forms.

THE QUANTUM ATOM
A *Nature* special issue
nature.com/bohr100

ILLUSTRATIONS BY THOMAS POROSTOCKY

Hollow atoms

The atom that Bohr proposed[1] in July 1913 looked like a miniature Solar System, with electrons arranged in concentric orbits around a positively charged nucleus. In Bohr's model, electrons were point-like particles that were quantized, meaning that they could jump from one orbit to another but could not exist in between. The advent of quantum mechanics in the 1920s retained the concept of orbits but re-imagined electrons as spreading everywhere around the nucleus. The location of each electron can be described only in probabilities, in the form of a mathematical wavefunction.

Electrons furthest from the nucleus can be kicked free with the least amount of added energy, so are usually the first to be stripped away. Yet X-rays, which pack a concentrated punch, can remove more tightly bound electrons from inner orbits. A medical X-ray takes out just one of those inner electrons before another from an outer shell drops down to fill the space. But the SLAC X-ray laser is in a class by itself. The beam is so intense and focused that every 100-femtosecond pulse sends 100,000 X-ray photons flying past each square ångström of space (1 ångström is 10^{-10} metres). That allowed Young to blast away all the inner electrons of the neon atoms in her 2009 experiment[2]. When electrons from the outer shells dropped into the abandoned inner shells, the beam soon kicked those out as well.

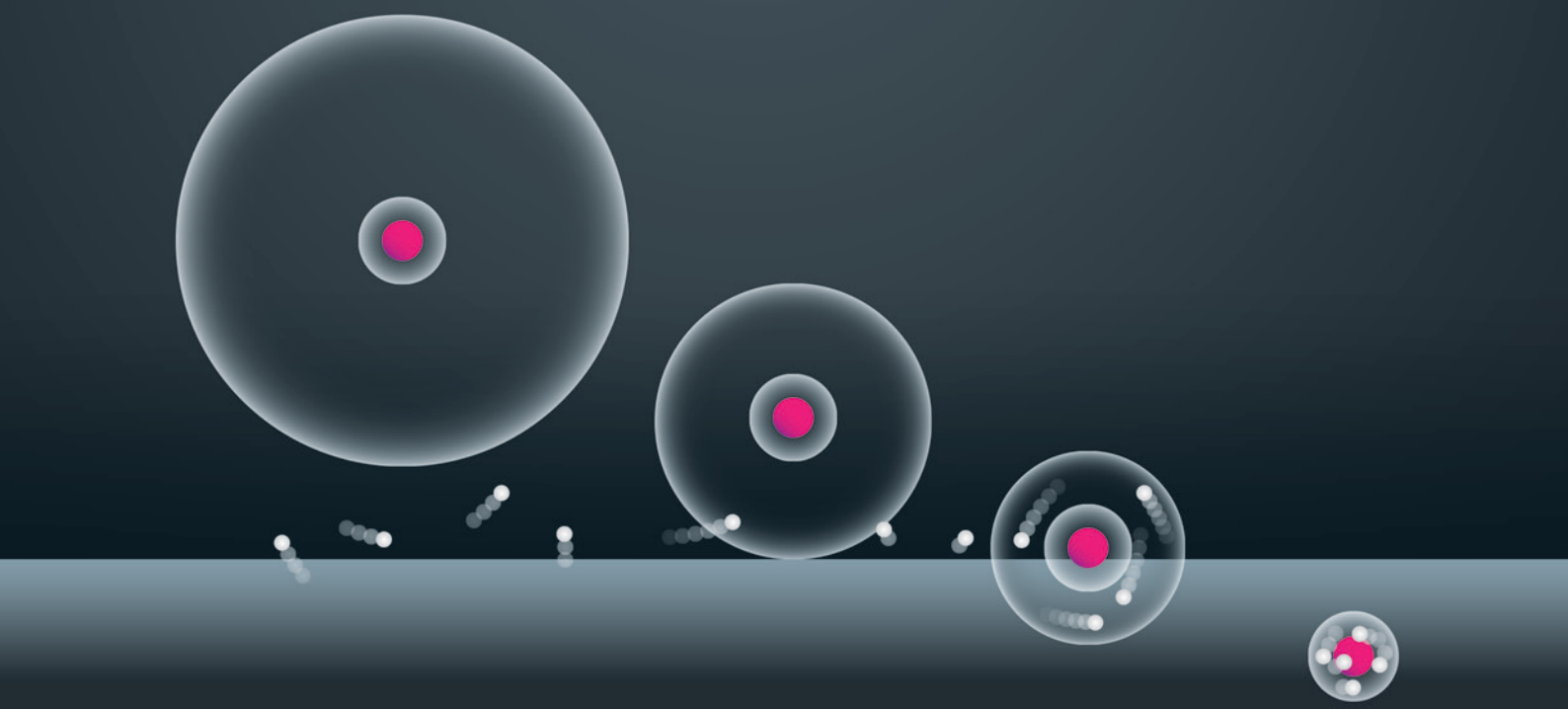
"If you tune your X-rays properly, you can pick which shell you want to empty out first," says Young. "Being able to control the inner-shell dynamics is very cool." The current record for this kind of atom-hollowing was reported last November[3] by a group at the Center for Free-electron Laser Science in Hamburg, Germany, which used the SLAC laser to strip away, from the inside out, the 36 inner electrons of a 54-electron-strong xenon atom.

Young hopes that research on hollow atoms will prove helpful when the laser is ready for one of its intended uses — creating images of biological molecules such as DNA and proteins by scattering X-rays off their atoms. Those pictures come at a price: the beam quickly destroys the molecules it is imaging. Knowing how hollow atoms form during this process may help researchers to interpret how the scattering pattern changes as a molecule explodes, Young says.

Two decades ago, several research groups made hollow atoms using a different process: first stripping almost all of the electrons from atoms, then depositing the resulting highly charged, slow-moving ions onto a surface. When the ions were a few tens of ångströms away from the surface, they attracted electrons from it, creating momentarily hollow atoms with electrons in outer but not inner shells. Those outer electrons then fell inwards, and the hollow atoms expelled a burst of energetic electrons and photons. "A hollow atom is nothing but a fireball of an enormous amount of energy," says Joachim Burgdörfer, a physicist at the Vienna University of Technology, who worked on developing the theory of the process[4].

Several research groups pursued hollow atoms in the late 1980s and 1990s, with some scientists exploring how the burst of photons from their formation might clean surfaces by removing the topmost layers without doing deeper damage. Although that procedure has been patented, it has not captured the attention of industry, says Fritz Aumayr, a physicist at the Vienna University of Technology. The closest it has come to an application so far was in 2008, when researchers invoked the process to explain how heavy ions spewed from the Sun can damage the surfaces of planets such as Mercury[5]. The ions become hollow atoms as they drop onto the planet, and release bursts of energy as they land.

This year, Aumayr published a paper[6] showing that the energy expelled from ions dropping onto carbon membranes can create nanoscale pores whose size is controlled by the strength of the ion's charge (that is, how many electrons it was missing). That might be a useful route for making nanosieves for filtering small molecules, he says, or for creating nanopores to pass DNA through for sequencing.



Giant atoms

From the perspective of an atomic nucleus, all electrons are far-flung voyagers. Whereas a nucleus measures femtometres in diameter, a bound electron typically travels 100,000 nuclear diameters away from the core. But Rydberg atoms, the colossi of the atomic world, have outer electrons so pumped with energy that they can travel 100 billion nuclear diameters — tens or hundreds of micrometres — from their nucleus. The largest Rydberg atoms even approach the size of the full stop at the end of this sentence.

Named after nineteenth-century Swedish physicist Johannes Rydberg, these giant atoms have been studied extensively since the 1970s, with the introduction of lasers that could excite electrons out to such vast distances. Like any distant traveller, the outer electron in a Rydberg system can be lonely and vulnerable. The attraction to the distant core is faint and easily disturbed by stray electromagnetic fields or collisions, so the atoms must be created in high vacuum. If carefully isolated from outside forces, such inflated atoms can be maintained for anything from a few hundredths of a second up to multiple seconds.

For Barry Dunning, a physicist at Rice University in Houston, Texas, the joy of Rydberg atoms is that they give physicists exquisite control over the motion of an electron. That is not possible with normal atoms because the electrons move much too quickly for even the fastest lasers. But the motion of an inflated electron in a Rydberg atom is much slower: it can be controlled with carefully directed nanosecond electric-field pulses, which allow researchers to herd the electron cloud by knocking it back and forth.

In 2008, researchers led by Dunning reported[7] that they had managed to squeeze the normally spread-out electron into a tight packet that briefly orbited the nucleus. Last year, they added radio waves that enabled that motion to be maintained indefinitely[8]. "It only took a century, but we recreated Bohr's atom," says Dunning proudly. His next idea is to try exciting and controlling two outer electrons at once, creating a system analogous to how Bohr might have pictured helium.

This kind of atom-stretching has some potential applications. Two gaseous atoms a few micrometres apart cannot normally affect each other. But inflate one (or both) to a Rydberg state, and the negatively charged electron clouds start to repel each other, distorting the energy levels of the atoms so that they are no longer isolated systems. Mark Saffman, a physicist at the University of Wisconsin-Madison, has used this property to make a quantum logic gate[9] — a fundamental part of a quantum computer — with lasers switching on a Rydberg interaction between two atomic quantum bits, or qubits.

He and other researchers hope next to add more atoms. A cloud of cold gas atoms might, if suitably excited, create a kind of hovering crystalline array of Rydberg interactions, says Matthew Jones, a physicist at Durham University, UK.

That approach might prove a useful model for studying the physics of 'strongly correlated' solid-state systems. These are systems, such as high-temperature superconductors, in which unusual properties emerge because particles interact strongly with their neighbours. An array of Rydberg atoms would not be a perfect model for the messy interactions in real solid-state systems, but the simplicity of the approach is a strength, says Burgdörfer. "It's a wonderful testing ground for probing many of these ideas about how strongly correlated physics actually works," he says.

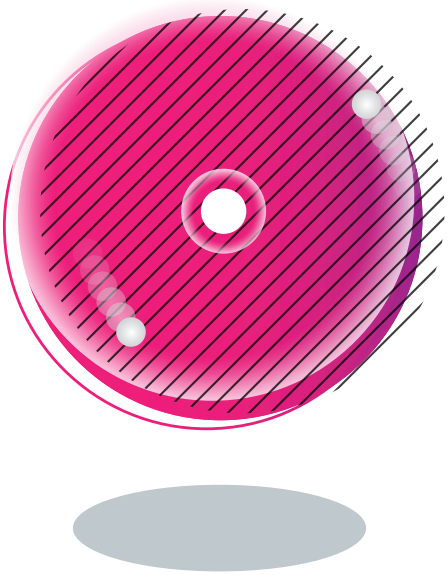
Antimatter atoms

The Large Hadron Collider at CERN, Europe's particle-physics lab near Geneva, Switzerland, currently lies in pieces, with engineers working on boosting its power. At the same time, in a side hall, an upgrade is taking place to an experiment that may allow physicists to measure the properties of atoms of antimatter.

It is a goal that researchers have been chasing since the first antihydrogen atoms were made at CERN in 1995. An antihydrogen atom consists of an antiproton and a positron, which respectively have the same mass as an ordinary proton and electron, but opposite charge. Beyond that, researchers know very little about antihydrogen. "Do matter and antimatter atoms obey the same laws of physics?" asks Jeffrey Hangst, spokesman for ALPHA, one of the collaborative efforts to make and analyse antihydrogen.

The experiments at CERN might also help to explain why there is more matter than antimatter in the visible Universe. The Big Bang should have created equal amounts of the two that would have annihilated on contact. But somehow, matter gained an advantage. Differences have been observed between the behaviour of some matter and antimatter particles, such as kaons and mesons, but these are far too small to explain the Big Bang conundrum.

To create antihydrogen atoms, researchers at CERN first make antiprotons by bombarding atoms with accelerated protons, then slow them down by passing them through metallic foil, cool them with cold electrons and trap them with electromagnetic fields. A similar trap accumulates positrons that are emitted by radioactive materials. When the clouds of charged particles are mixed, they make neutral antimatter atoms. But because these have no overall charge, in early experiments they easily escaped the electromagnetic fields



used to trap the charged antimatter particles.

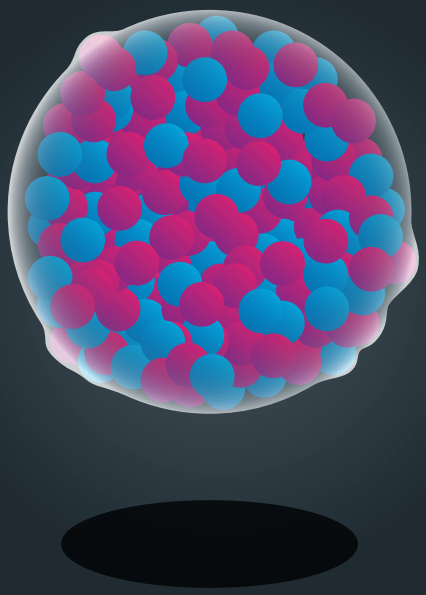
By 2002, two collaborations had been able to make as many as 50,000 atoms of antihydrogen, but the atoms quickly annihilated on the walls of their container. It took until 2010 before researchers at ALPHA showed[10] how to trap the atoms using three magnets with a combined field sufficient to restrain antihydrogen, with its tiny magnetic moment. At that time, the antimatter was held for just 170 milliseconds, and only about one atom was trapped for every eight times the group ran the 20–30 minute experiment, says Hangst. But the team has improved its equipment to trap one atom per experiment, and hold it for about 1,000 seconds.

ALPHA is now trying to probe the properties of the anti-atoms. This year, the team reported[11] watching the tracks of hundreds of antihydrogen atoms after they were released from their magnetic cage, to test whether antimatter falls up or down under gravity. The researchers do not yet have an answer, but the experiment works in principle, says Hangst. And in the upgrade, the team is moving in some lasers, with the idea of testing next year whether antihydrogen absorbs and emits light at the same frequencies as hydrogen.

Other teams at CERN are experimenting with different aspects of antimatter, such as how antihydrogen responds to changing magnetic fields. And researchers elsewhere are looking at even more exotic atoms: Ryugo Hayano, a physicist at the University of Tokyo, leads a team studying mixed matter–antimatter atoms, such as antiprotonic helium, in which a helium nucleus is surrounded by one electron and one negatively charged antiproton, an arrangement that lasts for a few microseconds.

In the end, such experiments may not find differences between matter and antimatter that are big enough to explain why the former has prevailed over the latter. But, says Hangst, "one never knows where the new physics might show up. You just have to keep looking."

Heavy atoms



Anti-atoms are rare, but researchers working with them are swimming in data compared with those chasing superheavy atoms. In an experiment that required prodigious patience, researchers at the GSI Helmholtz Centre for Heavy Ion Research in Darmstadt, Germany, spent almost five months last year firing titanium-50 ions — each with 22 protons and 28 neutrons — into a berkelium-249 target at the rate of about 5 trillion particles per second. The hope was that, just once or twice, two atoms would fuse to make an element with 119 protons, more than any created before.

Smashing beams of heavy atoms together has served physicists well over the past 70 years, allowing them to create increasingly heavy agglomerations of protons and neutrons, and to expand the periodic table far beyond the heaviest naturally occurring elements. The confirmed record-holder is element 116, livermorium, with 116 protons and, depending on the isotope, between 174 and 177 neutrons.

There have been claims to elements 117 and 118 too, but these have not been officially confirmed. And so far, "none of the current experiments have reported finding 119 or 120", says Christoph Düllmann, spokesman for the GSI-led collaboration — although he adds that his own team's analysis of last year's work is not quite complete.

There is a strong sense that the quest is coming to a dead end. The chance that nuclei will fuse decreases as they get heavier, because the protons and neutrons resist sticking together. Most researchers agree that beyond 120, the chance of getting two nuclei to fuse directly is vanishingly small. "So this leaves us with the question," says Düllmann, "what do we do next?"

To answer that requires an understanding of what motivates the superheavy search. Curiosity and national pride undoubtedly have a role, with politicians and scientists both looking to stamp their country's name into a new box on the periodic table. But each superheavy element is extremely short-lived, splintering in milliseconds.

Theorists have posited that some superheavy combinations of protons and neutrons will turn out to be stable for seconds, minutes or even days. This fabled 'island of stability' is thought to exist at between 114 and 126 protons, and around 184 neutrons. It is now clear that any attempt to make new superheavy elements by smashing a light particle into a heavier one will not reach the island: the isotopes spat out have too few neutrons. So researchers are changing tactics by trying to make heavier isotopes of elements that have already been created.

That is what scientists will attempt next year at the Joint Institute for Nuclear Research in Dubna, Russia. They plan to make neutron-rich isotopes of element 118 by firing beams of calcium-48 into radioactive californium-251.

The Russian team and others also want go back to the elements already made and create hundreds or thousands of atoms, rather than the handful necessary to claim a discovery. "We should set ourselves the goal of making not one or two atoms, but macroscopic quantities that we can use to study chemistry and nuclear structure in much greater detail," says Rolf-Dietmar Herzberg, a physicist at the University of Liverpool, UK. That might allow theorists to make more accurate predictions about where the island of stability lies.

But the temptation to expand the periodic table is strong. Researchers will probably turn away from head-on collisions and instead try knocking two heavy nuclei together in a glancing blow, which may stand a better chance of successfully fusing them to create new elements.

Physicists have a history of surprising themselves in their quest to create ever heavier atoms. In the early 1990s, no one thought that they could get past element 112 and then a tweak to the fusion process made it possible, says GSI team member Michael Block. "The next element is always the hardest."■

Richard Van Noorden *is a reporter for* Nature *in London.*

1. Bohr, N. *Phil. Mag.* **26,** 1–25 (1913).
2. Young, L. *et al. Nature* **466,** 56–61 (2009).
3. Rudek, B. *et al. Nature Photon.* **6,** 858–865 (2012).
4. Burgdörfer, J., Lerner, P. & Meyer, F. W. *Phys. Rev. A* **44,** 5674–5685 (1991).
5. Kallio, E. *et al. Planet. Space Sci.* **56,** 1506–1516 (2008).
6. Ritter, R. *et al. Appl. Phys. Lett.* **102,** 063112 (2013).
7. Mestayer, J. J. *et al. Phys. Rev. Lett.* **100,** 243004 (2008).
8. Wyker, B. *et al. Phys. Rev. Lett.* **108,** 043001 (2012).
9. Urban, E. *et al. Nature Phys.* **5,** 110–114 (2009).
10. Andresen, G. B. *et al. Nature* **468,** 673–676 (2010).
11. ALPHA Collaboration & Charman, A. E. *Nature Commun.* **4,** 1785 (2013).

COMMENT

UHLENBECK COLLECTION/AMERICAN INST. OF PHYSICS/SPL

Niels Bohr and his wife Margrethe around 1930.

The path to the quantum atom

John Heilbron describes the route that led Niels Bohr to quantize electron orbits a century ago.

In the autumn of 1911, the Danish physicist Niels Bohr set sail for a post-doctoral year in England inflamed with "all my stupid wild courage", as he expressed his state of mind in a letter to his fiancée, Margrethe Nørlund[1]. Bohr would need that courage on his route to his revolutionary quantum atom of 1913.

Bohr had reason to think himself designed for great things. He had won a gold medal from the Royal Danish Academy of Sciences in 1908, at the age of 23, for a theoretical and experimental study of water jets published by the Royal Society of London. His doctoral thesis on the electron theory of metals was so advanced that no one in Denmark could evaluate it fully.

Bohr went to the University of Cambridge, UK, to work with Joseph John (J. J.) Thomson, famous as the discoverer of the electron and recipient of the Nobel Prize in Physics for 1906. For Bohr, Thomson was "a genius who showed the way to everyone". But Thomson was too full of his own ideas to listen to those of a foreigner whose English ▶

NIELS BOHR ARCHIVE, COPENHAGEN

Niels Bohr (left) with Albert Einstein in 1925.

he had to struggle to understand. "They say he would walk away from the king," Niels wrote to his brother Harald, "which means more in England than in Denmark"[1].

Even if Thomson had been interested, he would have had trouble perceiving that his postdoc was a mature mathematical physicist. Furthermore, Bohr's speciality was criticism. In his thesis work, he had discovered errors in Thomson's papers, which he tried to bring to the professor's attention. That was not the right gambit.

Thomson was preoccupied with developing consequences of the model atom he had proposed in 1903. Later inappropriately and derisively nicknamed a 'plum pudding', it consisted of concentric rings of electrons rotating through a resistanceless spherical space that acted as if it were positively charged. In this picture, Thomson elucidated the periodic properties of the elements, the formation of simple molecules, radioactivity, the scattering of X-rays and β-particles, and the ratio between the weight of an atom and the number of its electrons.

Bohr spent much of his time at Cambridge attending talks and reading widely. He extolled Thomson's lectures and found much to admire in the treatise *Aether and Matter* (1900), in which Joseph Larmor, the occupant of the chair of mathematics once held by Isaac Newton, developed a world system based on electrons conceived as permanent twists in the ether. "When I read something that is so good and grand as that," Bohr wrote to Margrethe[1], "then I feel such courage and desire to try whether I too could accomplish a tiny bit."

NUCLEAR MODEL

In February 1912, Bohr went to the Victoria University of Manchester, UK, to arrange to work on radioactivity in Ernest Rutherford's laboratory. He looked forward to it in his understated way: "My courage is ablaze, so wildly, so wildly"[1]. Rutherford satisfied his expectations: "a really first-rate man and extremely capable, in many ways more able than Thomson, even though perhaps he is not as gifted"[1].

Rutherford certainly surpassed Thomson as a research director. When Bohr arrived, several men in the laboratory were working on implications of the nuclear model of the atom that Rutherford had introduced in 1911. To explain the unexpected reflection of α-particles from thin metal foils, detected by his research students, Rutherford had found it necessary to collect all the positive charge in Thomson's spheres into a tiny kernel at the atom's centre.

Soon Bohr joined in via his natural route: criticism. In calculating the transfer of energy from an α-particle to atomic electrons, Rutherford's theorist Charles Galton Darwin had not taken into account the resonance that occurs when the time of passage of the particle past the atom coincides with the natural frequency at which the perturbed electrons oscillate.

In improving the calculations, Bohr discovered that some modes of oscillation of a ring of electrons in the plane of their orbit grow until they tear the atom apart. This mechanical instability could not be mended by deploying accepted physical concepts. Bohr's thesis work had familiarized him with more general examples of failure in theories of heat radiation and magnetism that allowed electrons all the freedom that statistical mechanics granted them. To his unique way of thinking, the nuclear model appealed to Bohr precisely because it expressed this failure so conspicuously.

The model had further advantages. It made a clear distinction between radioactive and chemical phenomena, which in Bohr's view derived from the nucleus and the electronic structure, respectively. This inference was not as evident then as it is now. Even Rutherford had not yet grasped the distinction, and assigned the origin of β- and γ-rays to extra-nuclear electrons.

Most importantly, the nuclear model, combined with Rutherford's conception of the α-particle as a bare nucleus, almost thrust the concept of atomic number on physicists. They knew that the α-particle was a helium atom minus two electrons; its nucleus must therefore have a charge of two, implying that hydrogen's has a charge of one, lithium's a charge of three, and so on.

His confidence replenished by Rutherford's interest, Bohr drew up a memorandum in June or July 1912 to show how Max

Planck's idea that energy came in packets, or quanta, could extend the purview of the nuclear model to the problems that Thomson had considered, and to fix the size of atoms.

Although most of the memorandum was qualitative, in one essential point Bohr could be exact where Thomson could only estimate. Rutherford's scattering theory and experiments required that, for helium, the atomic weight (4) was twice the number of electrons (2). Thomson could only say, after extensive theoretical and experimental work on the scattering of X-rays and β-particles, that the number of electrons in an element was roughly three times its atomic weight.

After these first easy gains, Niels wrote to Harald, "Perhaps I have found out a little about the structure of atoms. If I should be right, it wouldn't be a suggestion of the nature of a possibility (i.e., impossibility as J. J. Thomson's theory) but perhaps a little bit of reality"[1].

Nonetheless, Bohr followed Thomson's lead in the other subjects he discussed with Rutherford: the periodic properties of the elements, determined by stability requirements imposed on their ring structures, and the binding of atoms into simple molecules, secured by exchanges of electrons.

To proceed with his calculations, Bohr laid down the ad-hoc postulate, conceived in analogy to Planck's radiation theory, that if the kinetic energy of each electron is proportional to the frequency of its orbit, it would neither radiate nor succumb to unstable oscillations, and he guessed that the constant of proportionality was a fraction of Planck's *h*.

BALMER'S NUMEROLOGY

Bohr's three-part paper on the constitution of atoms and molecules was published in the London-based *Philosophical Magazine* between July and November 1913. The second and third parts, which consider the periodic arrangements of the elements and molecular binding, record Bohr's debt to Thomson. Alone they would not have attracted attention or affected a revolution. What made Bohr's 'trilogy' memorable was its first part[2], on the spectrum of hydrogen, a subject he did not confront until February 1913.

A colleague asked him how he explained the formula for the frequencies of a series of spectral lines emitted by hydrogen, for which Johann Jakob Balmer had devised a simple arithmetical formula in 1885. Bohr replied that spectra were too complicated for his model, but had a look anyway. He saw immediately, so he later said, how to calculate the ratio of kinetic energy to orbital frequency for the model he had presented to Rutherford six months earlier (see 'Bohr's key to the microworld').

That early model had only a ground state, in which, by definition, the electrons have radiated away all the energy that nature allows them to dispose of. It could not explain why many frequencies are emitted. Bohr could see the bearing of the Balmer formula so quickly because, around New Year, he had extended his model in response to a remarkable series of papers by John William Nicholson, a mathematical physicist he had met in Cambridge.

Nicholson had matched the frequencies of many unattributed lines in solar and nebular spectra with the oscillations of electrons in a nuclear atom perpendicular to the plane of their ring. Unlike oscillations in the plane, perpendicular ones can be stable. By calculating the frequencies of rotation of his electrons from the spectra, he could compute their angular momenta. He found that, very closely, the angular momentum of each of his electrons was a small integral multiple of $h/2\pi$.

Nicholson's finding followed the lead of the *Conseil de Physique Solvay* of 1911, the conference at which Planck, Rutherford, Albert Einstein, Hendrik Antoon Lorentz and other luminaries considered problems in the theory of radiation. Discussion centred on Planck's concept of energy quanta: the simple harmonic oscillators by which he

THE BALMER FORMULA

Bohr's key to the microworld

The Balmer formula expresses the frequencies of some lines in the spectrum of hydrogen in simple algebra:

$$\nu_n = R(1/2^2 - 1/n^2)$$

where ν_n is the *n*th Balmer line and *R* is the universal Rydberg constant for frequency, named in honour of the Swedish spectroscopist Johannes Rydberg, who generalized Balmer's formula to apply to elements beyond hydrogen.

Following Max Planck's radiation theory, Niels Bohr converted the equation into units of energy, by multiplying both sides by Planck's constant, *h*. This allowed him to identify the energy of the electron in its *n*th state with the second term, $-Rh/n^2$. The first term would then be the negative of the energy for the second state ($n=2$), and the formula could be read to mean that a Balmer line originates in a jump of a hydrogen electron from its *n*th to its second state.

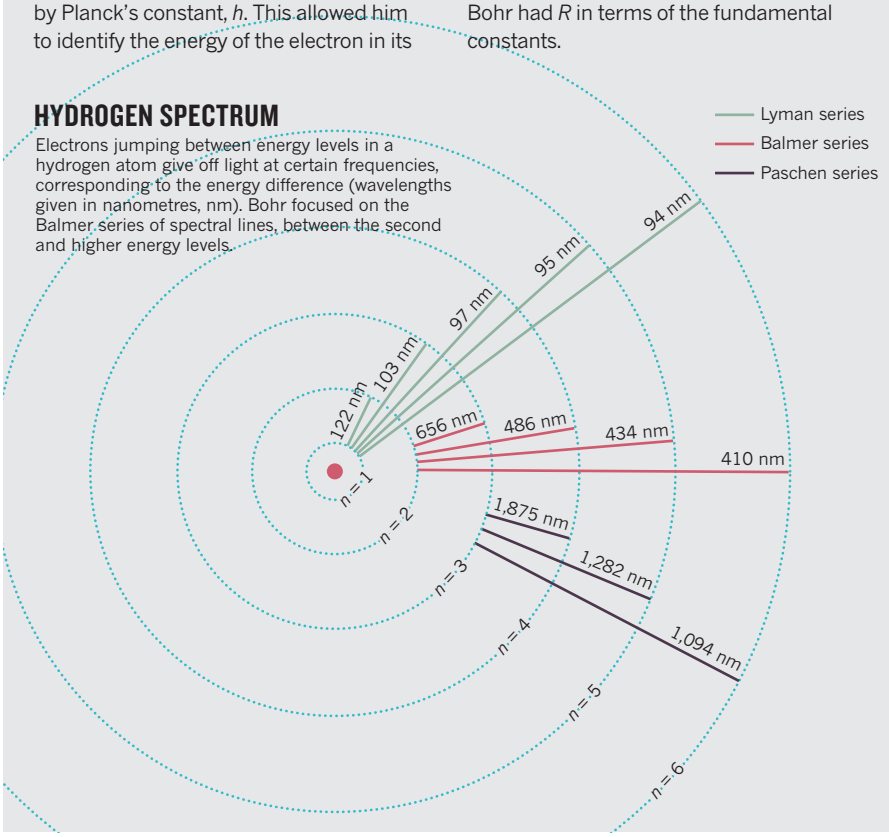
To calculate *R*, Bohr equated the energy of the *n*th state, $-Rh/n^2$, with the expression he already had for the kinetic energy T_n of an orbiting electron in his quantized model of the nuclear atom:

$T_n = 2\pi^2 me^4/h^2n^2$

where *e* and *m* are the charge and mass of the electron. Equating the two energies, Bohr had *R* in terms of the fundamental constants.

HYDROGEN SPECTRUM

Electrons jumping between energy levels in a hydrogen atom give off light at certain frequencies, corresponding to the energy difference (wavelengths given in nanometres, nm). Bohr focused on the Balmer series of spectral lines, between the second and higher energy levels.



represented material particles able to emit and absorb radiation possess energies only in integral multiples of their frequencies. An oscillator could emit or absorb radiation when its natural frequency equalled that of the radiation (ν), and then only in energy increments ($E = h\nu$), or quanta.

Because Nicholson's matches were astonishingly exact, and his model, like Bohr's, was both nuclear and quantized, Bohr had to take it seriously. Still reluctant to investigate spectra, he compromised, imagining that a captured electron occupied a sequence of excited states, radiating energy from each in Nicholson's manner as it descended towards the nucleus and the ground state.

Bohr introduced a running integer (n) into his model to handle the ascending scale of electron energies. By making the kinetic energy of the nth orbit proportional to n times the orbital frequency, Bohr easily obtained Nicholson's result about angular momentum and the additional information that the constant of proportionality was $h/2$. Thus Bohr had an integer-based series in his mind when he glimpsed the Balmer formula.

Using the relation $E = h\nu$, Bohr transformed arithmetic into physics by multiplying Balmer's formula by Planck's constant, h. Making it into an energy equation allowed Bohr to identify the kinetic energies of the various states with corresponding terms in the altered formula. That enabled him to derive the parameter in the Balmer formula, known as the Rydberg constant in terms of Planck's constant and the charge and mass of the electron.

The successful computation of the Rydberg constant demanded serious sacrifices from physicists. It made a Balmer line originate in a jump of an electron to the second orbit from a higher one, and put the explanation of such jumps beyond the reach of physics. Rutherford spotted this immediately: in order to 'vibrate' at the appropriate frequency, an electron would have to know where it would stop before it leapt. He was unwilling to concede foreknowledge to electrons or admit frequencies without vibrations.

Bohr replied that physicists must "renounce" — a word he came to use frequently — the possibility of exact descriptions of certain processes in the microworld.

Einstein perceived a greater loss. Planck had equated the frequencies of radiated light and mechanical oscillation. This was possible because the frequency of a simple harmonic oscillator is the same regardless of its energy. The oscillations of the radiator directly excited the 'ether', or the radiation field. But Bohr's jumps involved two orbits of different periods. The frequency of light emitted did not correspond with the motions of the electron supposed to produce it, contrary to the concepts by which physicists usually dealt with radiation.

NIELS BOHR ARCHIVE, COPENHAGEN

To develop his model, Bohr followed an analogy to the radiation theory of Max Planck (right).

Bohr's sense of responsibility directed him to attempt to anchor the basic postulate of his quantum atom — that the ratio of kinetic energy to orbital frequency in the nth state is proportional to $nh/2$ — in deeper foundations. He did not find the job easy. The first instalment of the trilogy contains four distinct, and largely contradictory, attempts.

Two of them develop the analogy to Planck's radiation theory that provided the form of Bohr's postulate. The third foundation is altogether different. It requires that in jumps between very large neighbouring orbits, where the electron is almost free from the nucleus, the radiation frequency is asymptotically equal to the frequency of the orbits, which are asymptotically equal to one another. This anticipates Bohr's correspondence principle, according to which, at an appropriate limit, calculations of a physical quantity must give the same numerical result in ordinary physics and in quantum theory.

By the end of 1913, Bohr had given up the Planck pedigree as "misleading" (the nuclear atom is not a simple harmonic oscillator) and adopted the correspondence principle as his preferred foundation. He also retained the fourth formulation, the only one now remembered: the quantization of the angular momentum (which follows from the basic postulate by replacing the ratio of kinetic energy to orbital frequency by its mechanical equivalent, π times angular momentum). As a condition on the orbit, the fourth foundation differs conceptually from the other three, which relate the orbit to the radiation emitted by an electron undergoing a quantum jump.

OPEN TO AMBIGUITY

Bohr's ability to entertain several conflicting ideas, and his courage in demanding sacrifices of physicists like Einstein, Planck and Lorentz, are breathtaking. We know that he did not lack confidence. Blazing courage is one thing, but tolerance of ambiguity is quite another.

Correspondence with his immediate family, especially Margrethe, suggests sources for this tolerance. Well before he became entangled in the quantum atom, Bohr had developed a doctrine of multiple partial truths, each of which contained some bit of reality, and all of which together might exhaust it. "There exist so many different truths," he wrote to Margrethe. "I can almost call it my religion, that I think that everything that is of value is true."[1]

Bohr's seminal analysis of the Balmer spectrum expressed the partial truth of Planck's radiation theory and the partial truth of classical physics. Bohr may have owed his notion of partial truth at least in part to ideas he found in the writings of his professor of philosophy, Harald Høffding, and in William James's *Pragmatism*, published in 1907, which Bohr may have known from Høffding.

Family letters hitherto unavailable, which will be published in part next month by Finn Aaserud and myself in *Love, Literature, and the Quantum Atom*[1], open new directions in which to explore this connection, about which historians and philosophers have speculated on the basis of the later and slighter evidence of Bohr's principle of complementarity.

Between periods in which his courage blazed and his blood boiled, Bohr was subject to the sorts of self-doubt that ordinary people have. As their correspondence shows, Margrethe played an important, perhaps an essential, part in smoothing out Niels' mood swings and reassuring him that he was the great man his Danish support system took him to be.

In many letters he asks her to help him pay his debts, by which he meant the obligations he felt he owed for his great gifts, for the encouragement he had received to develop them and, perhaps, for the wider perspectives he gained in England. He could discharge these debts only by great deeds. He made a huge down payment to these imaginary creditors — including Thomson and Rutherford — with his revolutionary quantum atom of 1913. ■

John L. Heilbron *is emeritus professor of history at the University of California, Berkeley, USA.*
e-mail: john@heilbron.eclipse.co.uk

1. Aaserud, F. & Heilbron, J. L. *Love, Literature, and the Quantum Atom: Niels Bohr's 1913 Trilogy Revisited* (Oxford University Press, 2013).
2. Bohr, N. *Philos. Mag.* **26,** 1–25 (1913).

DAN MCCOY/RAINBOW/SCIENCE FACTION/CORBIS

Neutrinos are among the many types of particles produced when electrons and positrons collide.

The enigmatic electron

While we are still learning about the particle's true nature, says **Frank Wilczek**, let's celebrate its beauty.

What is an electron? That question was central to the development of quantum theory early in the twentieth century, and remains at the frontier of physics today. There are several inconsistent answers, each correct. A century after Danish physicist Niels Bohr conceived of the electron as the proton's satellite[1], our perception of the electron continues to evolve and expand.

Bohr's answer to this question in 1927 epitomized his beloved concept of complementarity: in some circumstances electrons are best described as particles, with definite positions; in others as waves, with definite momenta[2]. Both descriptions are valid and useful, yet according to Heisenberg's uncertainty principle, they are mutually exclusive: positions and momenta cannot be known accurately at the same time. Each depiction captures an aspect of the electron's nature, but neither exhausts it.

Modern quantum theory reinforces Bohr's conclusion that what you see depends on how you choose to look. Electrons are both ideally simple and unimaginably complex. They are understood with precision yet remain mysterious. Electrons are stable bedrock in physicists' world picture, and are playthings that we are learning to fragment and transform.

SIMPLE AND COMPLEX

For most practical purposes, an electron is a structureless particle with an intrinsic angular momentum, or spin. Just two numbers — the electron's mass and its electric charge — fuel the equations that describe its behaviour. From this 'practical electron' model, physicists constructed modern microelectronics. It is also the working foundation for chemistry, including biochemistry.

But to a high-energy positron (anti-electron), an electron is a cornucopia. Collisions of electrons and positrons, such as those carried out at the Large Electron–Positron (LEP) collider at CERN, Europe's particle-physics lab near Geneva in Switzerland, produce streams of quarks, gluons, muons, tau leptons, photons and neutrinos. To understand the complexity of an electron, all of the esoteric resources of modern physics must be brought to bear.

There is tension between these two observations, that the electron is a simple point-particle, and that it contains the world. They can be reconciled through a concept that I call quantum censorship, whereby properties of objects vary according to the energy with which they are probed. Quantum censorship was implicit in Bohr's atomic model and, in a more general form, remains a central pillar of modern quantum theory.

In his 1913 model of the hydrogen atom[1], Bohr pictured an electron orbiting the proton like a planet in a miniature Solar System. As he knew, and the physicist James Clerk Maxwell had emphasized before him, mechanical models of the atom have severe problems. They predict a variety of hydrogen atoms, with different orbital shapes and sizes, whereas in reality, all hydrogen atoms are identical. The models also predict that atoms are unstable, because moving electrons should radiate energy and spiral into the central proton, which clearly they do not.

Bohr boldly assumed away those difficulties. He restricted electrons to a set of discrete, or quantized, energy states within an atom to avoid instability. He recognized that the level with the lowest energy, or ground state, has a finite size, keeping the electron and proton apart.

Today, we trace Bohr's rules to the fact that the proper quantum-mechanical description of electrons involves wave functions, the oscillation patterns of which are standing waves. The equations that govern electrons in atoms are similar to those for vibrations in musical instruments, which produce scales of distinct tones.

The same ideas apply to complex, bound systems, such as atoms that have many electrons and larger nuclei. A system in its ground state tends to remain there, if little energy is fed in, betraying no evidence of its internal structure. Only when it is excited into a higher state do complexities emerge. This is the essence of quantum censorship. Thus, below an energy threshold, atoms

appear to be the "hard, massy, impenetrable" units that Isaac Newton inferred. Above it, their components can be torn out.

Similarly, electrons, despite the fecundity that they showed at the LEP collider, betray nothing of their inner workings at low energies. An electron's structure is revealed only when one supplies enough energy to unleash electron–positron pairs — at least 1 megaelectronvolt, which corresponds to the unearthly temperature of 10^{10} kelvin. Thus the practical electron is not an approximation to reality, in the usual sense of fuzziness; rather, it is a precise description that applies under limited (albeit quite generous) conditions.

Having recognized its power, let us celebrate the practical electron's intellectual beauty. Each of its properties is intimately connected to profound symmetries of physical law. Mass and spin classify all possible realizations of special relativity by particles. Electric charge, a conserved quantity, classifies realizations of the 'gauge symmetry' of electromagnetism. Specifying how the practical electron responds to those symmetry transformations determines its physical behaviour. The electron is thus an embodiment of symmetry: its physical properties are inherent to its mathematical form.

PRECISE AND MYSTERIOUS

In principle, electrons can possess both magnetic- and electric-dipole fields, the axes of which are set by the electron's spin. But the status of these fields could hardly be more different. The strength of the electron's magnetic field provides perhaps the most stringent and brilliantly successful comparison of theory and experiment in all of physical science, whereas the value of the electric field has never been measured. It is a mystery even to theory.

Establishing the strength of the electron's magnetic field — in terms of a gyromagnetic ratio or '*g*-factor' — was a major focus of twentieth-century physics. An early triumph of physicist Paul Dirac's 1928 relativistic wave equation for the electron[3] was its suggestion that $g = 2$, which was found to be nicely consistent with atomic spectroscopy.

Post-war developments in precision spectroscopy, using atomic beams, revealed that *g* deviated from that value by one part in 1,000. Theorists matched that deviation when they had mastered the mathematical difficulties of quantum-field theory enough to calculate corrections to the Dirac equation to account for quantum fluctuations (the energy of which release virtual photons).

Creative dialogue between experiment and theory continues today, with improved accuracy on each side allowing ever more rigorous comparisons. The experimental frontier has moved to beautiful investigations of single electrons in electric and magnetic traps. Theoretical calculations have become intricate, now including fluctuations in fluctuations in fluctuations. The value of *g* is known to a dozen significant digits[4].

A crude but appealing 'explanation' of the origin of the electron's magnetic field is that quantum uncertainty in position smears the electron's charge over a volume, which rotates because of the electron's spin. The electron is effectively a spinning ball of charge, and elementary electromagnetism tells us that this generates a magnetic-dipole field. The size of that ball can be estimated to be roughly 2.4×10^{-12} metres. Attempts to pin down an electron's position more accurately than this require, according to the uncertainty principle, injecting the electron with so much energy that extra electrons and anti-electrons are produced, confusing the identity of the original electron.

"An electron's structure is revealed only when one supplies enough energy."

An electric dipole, should it exist, would generate broadly similar corrections. But no such field has been detected. Great efforts have gone into the experimental search, using all the tricks and traps that revealed the magnetic moment. So far there is only an upper bound for the electric dipole moment[5]. This is an extraordinary 17 orders of magnitude smaller than one might expect — naively, given the electron's effective size.

Why is it so hard for spin to align electric charge? One explanation involves time-reversal symmetry. If we run time backwards, the laws of physics stay the same. But for a spinning electron, the north and south poles would swap. Thus an electric dipole accumulating charge at one pole violates time-reversal symmetry.

But nature does not always respect time-reversal symmetry, as we know from observations of *K* and *B* mesons[6]. So a non-zero electric dipole moment for electrons is a theoretical possibility. It is tantalizing that values of the electric field that lie just below the present upper bound are expected in many theories of physics beyond the standard model of particle physics, including supersymmetry. Ingenious experiments using solid-state physics and molecular spectroscopy have been proposed in a bid to search more sensitively for the existence of tiny electric fields that are generated by re-orienting spins. This 'other' dipole moment might prove to be a focus for twenty-first-century physics.

RIGID AND PROTEAN

Electrons are rigid and defend their integrity stoutly. They follow the Pauli exclusion principle, which states that no two electrons can be in the same quantum state at the same time. This is the defining characteristic of fermions, a class of particles that includes protons, neutrons and electrons. As a result, electrons cannot be crushed.

Nature's most imposing macroelectronic creation is the white-dwarf star. The Sun will become such a star 4 billion to 5 billion years from now, when it has exhausted its nuclear fuel, causing it to collapse into a sphere roughly the size of Earth, but a million times more dense. White dwarfs rely on the quantum statistics of electrons for their support. Squeezing electrons together promotes some into higher energy states, exerting a force or 'degeneracy pressure' that balances gravity and halts further collapse.

But subtle collective action can achieve what raw pressure does not, and fragment electrons. This has been extensively studied in electron states in thin semiconductor interfaces that are extremely pure and cold, and are subjected to strong magnetic fields. These states are known as fractional quantum Hall effect liquids[7]. Their electric currents reveal the presence of particles whose charge is a fraction of an electron's.

Electrons also lose their individual identities in superconductors, in which electrons pair up to form a pervasive sea. Thus, electrons become their own antiparticles. By combining fragmentation with superconductivity, we can get half-electrons that are their own antiparticles. Such 'Majorana modes' have now been observed experimentally[8] and promise to have exotic properties. Notably, their quantum state retains 'memories' of how they were created and where they have been. Manipulating electron fragments opens up rich new possibilities for microelectronics and quantum computing, which are only beginning to be explored.

So, what is an electron? An electron is a particle and a wave; it is ideally simple and unimaginably complex; it is precisely understood and utterly mysterious; it is rigid and subject to creative disassembly. No single answer does justice to reality. ■

Frank Wilczek *is professor of physics at the Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*
e-mail: wilczek@mit.edu

1. Bohr, N. *Philos. Mag.* **26,** 1–25 (1913).
2. Bohr, N. *Nature* **121,** 580–590 (1928).
3. Dirac, P. A. M. *Proc. R. Soc. A* **117,** 610–624 (1928).
4. Mohr, P. J., Taylor, B. N. & Newell, D. B. *Rev. Mod. Phys.* **84,** 1527–1605 (2012).
5. Hudson, J. J. *et al. Nature* **473,** 493–496 (2011).
6. Kirkby, D. & Nir, Y. 'CP violation in meson decays' in Beringer, J. *et al.* (Particle Data Group) *Phys. Rev. D* **86,** 010001 (2012).
7. Das Sarma, S., Pinczuk, A. (eds) *Perspectives in Quantum Hall Effects* (Wiley-VCH, 1996).
8. Mourik, V. *et al. Science* **336,** 1003–1007 (2012).

A sperm whale (*Physeter macrocephalus*) and a snorkeller in Caribbean waters.

JONATHAN BIRD/GETTY

NATURE WRITING

Cetacean subtleties

Callum Roberts enjoys a celebration of the oceans and their largest denizens.

In December 2010, a whale and her calf beached and died on a remote stretch of New Zealand's Bay of Plenty. Two years later, the find was announced to the world as the first sighting of the planet's rarest cetacean. *Mesoplodon traversii*, the spade-toothed beaked whale, was previously known only from three skull fragments collected over 140 years from islands scattered across the South Pacific.

This whale is just one of many remarkable creatures that Philip Hoare describes in *The Sea Inside* — a beautifully written, impressionistic memoir of a seagoing life. The book is threaded through with ornithology, cetacean science, literature and a host of figures, from eighteenth-century medic and menagerie-keeper John Hunter to the eccentrics Hoare meets on his travels from England to New Zealand. Looking at a tide-worn whale skull in the storeroom of the Te Papa museum in Wellington, Hoare imagines the creature alive in its ocean home, "striated and crisscrossed by innumerable scratches, as if subjected to cosmic strikes".

For me, nothing underlines the immensity of the world's oceans quite like this: that an animal roughly the size of a hippopotamus, surfacing to breathe many times a day, could have lived unseen for so long. But although the vastness of the oceans is central to planetary processes, it is no mark of invincibility. The human imprint can be found at the bottom of even the deepest abyss, where plastic bottles and tin cans lie incongruously alongside life forms so strange that they seem almost alien.

The Sea Inside
PHILIP HOARE
Fourth Estate: 2013. 384 pp. £18.99

Hoare is acutely aware of this vulnerability and worries about where the oceans and their inhabitants, many from ancient evolutionary lineages, are headed. His greatest love is cetaceans (he is author of the prizewinning *Leviathan or, The Whale;* Fourth Estate, 2008) and he describes the extraordinary lengths to which he has gone to swim with them in the wild: blue whales in Sri Lanka, Hector's dolphins in New Zealand, and Risso's dolphins and sperm whales in the Azores. These are vivid encounters. After days hanging at the surface in water more than 3,000 metres deep with a group of sperm whales, he learns to read their behaviour by subtle signs, just as "you can see in a person's eyes what they think of you long before they might put it into words".

But Hoare acknowledges how hard it is to know what a whale thinks, because its senses are tuned differently from ours. We are hugely dependent on vision, whereas a whale relies far more on hearing in its underwater environment. Hoare contends that whales' barrage of acoustic clicks allows them to probe solid bodies that have a similar density to water, much as a sonogram images a fetus. In a comment that floored me, he asserts that "a whale or dolphin can see the interior of my body as accurately as I can see the exterior of hers". Although plausible, I have not seen research that backs this up. It would be wonderful if it were true.

Slowly and unwittingly, humans have pushed some of the most spectacular and iconic species to the edge of extinction. Although a few still hover on the brink, many whale species — from humpback to bowhead — are making a comeback after decades of protection. Nobody celebrates the joy of their resurgence better than Hoare. ■

Callum Roberts *is professor of marine conservation at the University of York, UK, and author of* Ocean of Life: How our Seas are Changing.
e-mail: callum.roberts@york.ac.uk

BIOTECHNOLOGY

Genomics and us

Michael Rawlins examines a call for biotechnology to be geared towards public health.

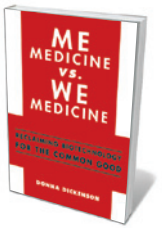
Personalized medicine has been a major ambition of clinical pharmacology for more than 40 years. The mantra of giving 'the right drug to the right patient, and at the right dose' has been accepted as the goal in seeking to maximize a drug's effectiveness and minimize its toxicity.

There have been modest successes. For example, in the 1970s researchers came up with the idea of using simple measurements of different patients' renal functions to tailor doses of drugs predominantly excreted unchanged in the urine — such as digoxin, the cardiac glycoside. And for the past three decades, the oestrogen antagonist and breast cancer treatment tamoxifen has been targeted solely at women with oestrogen-receptor-positive tumours.

Recent advances in genetics and genomics have shown that much more may be possible. We now know, for example, that there are important associations between mutations in human leukocyte antigen genes and the development of severe — sometimes lethal — adverse reactions to drugs such as abacavir (for HIV and AIDS) and carbamazepine (for epilepsy). It is now possible to identify patients who should avoid these drugs. Safe and effective warfarin doses can also be more reliably predicted by genotyping the enzymes involved in the drug's metabolism. And contemporary molecular genetics has become an important tool for monitoring the spread of infectious diseases and determining the antigenic components of influenza vaccines.

In *Me Medicine vs. We Medicine*, however, Donna Dickenson argues that the focus on personalized (me) medicine is eclipsing the focus on public health (we medicine). She is concerned that the personal-genomics revolution has yet to live up to the hype and that simple measures designed to maintain good health and prevent illness are being squeezed out by public and private funders, researchers, companies and health providers.

Dickenson is at her best when discussing the benefits of immunization in the context of public health. Taking a historical and sociological perspective, she provides a useful ethical analysis of the importance of herd immunity and the limitations of arguments that reject the concept of individual freedom among populations. She writes about the "free riders" who rely on herd immunity to avoid vaccination. She also reminds readers of mandatory US and UK vaccination programmes. In nineteenth-century England and Wales, for instance, the Poor Law Guardians were authorized to seek out and discipline non-compliers. Dickenson does not support the "vaccine sceptics", but places their concerns in this historical context.



Me Medicine vs. We Medicine: Reclaiming Biotechnology for the Common Good
DONNA DICKENSON
Columbia University Press: 2013. 304 pp. £19.95, $29.95

It is in her foray into personalized medicine that I feel her arguments become unsustainable. Dickenson writes that she has undertaken "a reality check" on the claims made for personalized medicine, and asserts that the evidence fails to support them. She also says that resources are preferentially given to personalized medicine, endangering public-health measures such as immunization. However, she provides no details on how she reviewed the literature or analysed the results.

Dickenson castigates Francis Collins, the geneticist who led the Human Genome Project, for describing personalized medicine as a "paradigm shift" in the way clinicians will be able to practise in future. She criticizes that project as "very generously funded, without having so far produced correspondingly weighty results for translational medicine".

She is not, of course, the only person to voice this opinion — Collins himself admitted in this journal that "those who somehow expected dramatic results overnight may be disappointed". But translating basic medicine into the clinic takes time. More than 40 years elapsed between the acceptance of the 'germ theory' of disease and the introduction of Salvarsan to treat syphilis; and 60 years passed between German physician Robert Koch's identification of the cause of tuberculosis and the discovery of streptomycin.

Dickenson devotes several chapters to the failings she perceives in direct-to-consumer genetic testing (such as that carried out by personal genetics company 23andMe, based in Mountain View, California) and private-cord-blood banking, which involves storing a baby's stem-cell-rich umbilical cord for its family's future medical use. However, in my view she does not fully explain how these procedures might have an adverse influence on personalized medicine.

Dickenson says that such 'retail' genetic analysis results in overwhelming additional demands on health-care systems. This is not necessarily true. For example, 23andMe tests for pseudocholinesterase deficiency, which affects about 1 in 1,000 people in some populations. The deficiency reduces the body's ability to metabolize some muscle-relaxant drugs used in anaesthesia, which can trigger temporary respiratory paralysis. Prior knowledge of a person's pseudocholinesterase status means an alternative agent can be used and resources can be saved.

Dickenson believes that for-profit commercial organizations largely support the development of personalized medicine. She cites the case of Herceptin (the brand name of the trastuzumab antibody), which is effective in the one-third of women with breast cancer whose tumours express the human epidermal growth factor receptor 2 (HER2) protein. Limiting use of the medicine to such women prevents use in those who would experience only its adverse effects, a particular concern given its potential to cause heart failure. However, because the manufacturer — Genentech of San Francisco, California — also holds the patent on the *HER2* gene, Dickenson says that other organizations are prevented from developing their own *HER2* antagonists. But this is not so — global health-care company GlaxoSmithKline has developed and marketed its own *HER2*-receptor antagonist, lapatinib (brand name Tykerb).

She also writes that Herceptin was only introduced in the UK National Health Service in response to public pressure. In fact, this happened as a result of guidance produced by London's National Institute for Health and Care Excellence during my time there as chairman, after taking careful account of its clinical- and cost-effectiveness.

Personalized medicine and public health are complementary approaches to maintaining the health of populations. And as most public-health measures are either cost-saving or cost-neutral in the long term, there is no cost-effectiveness conflict with personalized medicine. However effective our public-health measures, we all become sick. And when we do, the personalized medicine of the future offers great promise. ■

➲ NATURE.COM
For *Nature*'s Human Genome at Ten special, see:
nature.com/humangenome

Michael Rawlins *is president of the Royal Society of Medicine, London. The author was chairman of the National Institute for Health and Care Excellence from 1999 to 2013.*
e-mail: president@rsm.ac.uk

FAYAZ KABLI/REUTERS

Changing up: a Kashmiri money changer charges a small fee to swap damaged money for new notes.

ECONOMICS

A tale of cash and credit

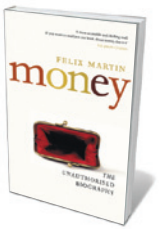
Martin Shubik navigates a study of the complex web of relationships that gives money meaning.

The macroeconomist Felix Martin covers a vast geographical and historical spread in his argument that we have it all wrong about money. *Money: The Unauthorised Biography* reveals that credit has a crucial role in society, but that many misunderstandings persist about the relationship of credit to gold or fiat money. The latest views on microeconomic theory get no airing here, yet such developments point to a scientific shift towards an information-processing and network analysis of money and credit markets in a dynamic evolving economy.

To a degree, Martin shows, the concept of money is all in the mind. Trust effectively oils the wheels, and a complex blend of expectations, law, customs and force creates money's 'store of value', such as its symbolic role representing gold. The key element is that transferable credit serves as money if the issuer (the government, banks, corporations or individuals) is perceived as creditworthy. The fundamental property of both money and credit in the dynamics of trade is the expectation that the recipient will be able to pass payments received on to others without loss.

Martin cites the example of the stone money of Yap, a Micronesian island in the western Pacific. The island's traditional monetary system is based on calcite discs (some of them 4 metres in diameter), nicely illustrating how geographical isolation and a tightly knit and networked society allow for the acceptance and enforcement of a credit system in which ownership rights are reassigned to a group of otherwise useless stones.

The author points to Ireland's banking system as another example of trust in a highly networked society. The country's economy survived reasonably well when strikes closed its banks for half a year in 1970. Among Ireland's relatively sedentary and homogeneous population, trade was carried out with cheques, aided in part by the informal credit-evaluation system provided by customers of pubs.



Money: The Unauthorised Biography
FELIX MARTIN
Bodley Head: 2013. 336 pp. £20

Martin then gives a swift overview of the evolution of trade throughout history, noting that the bureaucracy and the accounting and bookkeeping system developed by the ancient Mesopotamian civilization in the eighteenth century BC and earlier helped to simplify and standardize trade. Here he brings up the central problem of all monetary economies: who controls the system?

As we know from Aristotle, the Greeks saw money as facilitating individual exchange; conversely, the early Chinese viewed it as a key tool of the state. Martin sketches the growth of monetary sophistication in Europe, with regular trade fairs, bills of exchange and clearing houses providing the sinews for international trade and setting the stage for the emergence of central banks. He flags up the problem of debt, and explains the ways in which the public and politicians misunderstand the roles of money, credit and the uses and dangers of debt. He also notes the importance of financial inventions such as the automated teller machine (ATM), and the dangers of misconceiving the political and bureaucratic sophistication and support needed to preserve public confidence in new methods of finance.

Switching to the subject of the recent financial upheavals, Martin cites the views of economists Walter Bagehot and Hyman Minsky (as would I) in reference to crises in highly complex credit-based economies. Bagehot, the renowned one-time editor of *The Economist*, also authored the enduring masterpiece *Lombard Street: A Description of the Money Market* (1873). Minsky was an important twentieth-century US economist who worked on the subject of financial instability. In keeping with the advice of these figures, Martin notes the importance of central banks in making bold and decisive moves in times of crisis — but his analysis is much too brief. It is difficult to convey the degree of financial plumbing and public trust needed in an honest, lean and efficient bureaucracy to move from talk and promises to implementation.

The book's main weakness is its failure to cover many relevant developments in the microeconomic theory of pricing systems, decentralization, information and the development of trust. There is no reference to the vast literature on agency theory, a concept that concerns asymmetric information (A knows things that B does not know and can lie to him) and that makes credit assessment and bureaucratic behaviour so opaque. No reference is made to contract theory or mechanism design, devoted to designing and testing new structures that incorporate appropriate incentive systems in economic institutions such as commodity markets, clearing houses and web-banking services. Nor is there any appreciation of the considerable developments in game theory or work in the fields of econophysics and bioeconomics.

Martin does, however, understand that credit is central to running a modern economy and that it is a delicate flower that must be nurtured by both government and private financial institutions. He takes money seriously, but not the sea change in basic economic theory aimed at understanding the evolution and control of trust. ■

Martin Shubik *is the Seymour Knox Chair Professor of Mathematical Institutional Economics (emeritus) at Yale University, Connecticut. His books include* The Theory of Money and Financial Institutions.
e-mail: martin.shubik@yale.edu

NORBERT WU/MINDEN PICTURES/CORBIS

An ice-breaker ship cuts a route through ice floes in McMurdo Sound, Antarctica.

ENVIRONMENT

Piscine plunder

Michael White assesses a film documenting the exploitation of Antarctica's pristine Ross Sea.

In the 1999 science-fiction blockbuster film *The Matrix*, Agent Smith labels *Homo sapiens* a "cancer of this planet", declaring "you multiply and multiply until every natural resource is consumed and the only way you can survive is to spread to another area." Although Peter Young does not go that far in his glorious but flawed documentary *The Last Ocean*, the film is a powerful statement about humanity's urge to chase resources over every inch of Earth.

Young's focus is the case to end fishing of the Antarctic toothfish (*Dissostichus mawsoni*) in the pristine Ross Sea, a deep Antarctic bay. Marketed as Chilean sea bass, the toothfish is delicious and versatile — virtues that are also its vulnerabilities. A relative, the Patagonian toothfish (*Dissostichus eleginoides*), has been over-exploited in much of the Southern Ocean (peaking at a reported catch of 40,000 tonnes in 1995) and this has increasingly pushed fishing into the Antarctic waters where *D. mawsoni* lives.

Young builds his conservation argument around a few main themes, one of which is unquestionably valid and others less so. The primary focus is that the Ross Sea brims with an astonishing diversity of life. The message is brought home with contemplative and elegantly sparse cinematography and audio recording: a lone penguin races across the ice; a Weddell seal emits an otherworldly call. Young's argument that there is nowhere else like the Ross Sea left on the planet is sound and is backed by the many participating scientists, including Antarctic ecologist David Ainley.

The Last Ocean
DIRECTED BY PETER YOUNG

But there is a whiff of conspiracy theory about *The Last Ocean*. Young suggests that industry funding seriously undermines the credibility of the Marine Stewardship Council — a global certifier of sustainable fisheries, including those that support the toothfish. However, similar certification practices are common in many industries. Young also says that the New Zealand fishing industry, with government support, mounted a coordinated and unjustifiably defensive response to a journalistic critique of the toothfish industry published in July 2010 — yet the industry's response in the film comes across as bumbling.

Other arguments are even murkier. New Zealand government officials are portrayed as callous and exploitative, but the film fails to adequately acknowledge the importance of the Ross Sea toothfish fishery in cementing New Zealand's claim to the Ross Dependency, a huge slice of ocean and bits of land ceded by Britain in 1923.

The fishing industry is expected to reduce existing toothfish stocks by 50% over the next 35 years, a limit set by the regulatory Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR), which grew out of the 1961 Antarctic Treaty. A central scientific argument against extraction down to such levels is that sustainability cannot be guaranteed because we lack sufficient scientific knowledge about this species. Although this is probably true, an industry representative in the film points out that there would be little fishing of any kind if such guidelines were widely imposed.

But the facts stand. The toothfish is not fished for subsistence, and this pleasure-seeking has brought clear effects in less than 20 years. The effort per unit of fish caught is increasing, the Ross Sea killer whales that count on the toothfish as a main food source are disappearing and the decades-long toothfish research programme in McMurdo Sound has gone — along with the sound's population of *D. mawsoni*.

Achieving a fishing ban seems unlikely, as does the greater goal of protecting the Ross Sea. Decisions about fisheries around Antarctica are regulated by the CCAMLR. Regulations are changed by consensus, a system that is subject to endless political bickering. Near the end of the film, the United States and New Zealand introduce competing plans for marine protected areas designed to preserve huge chunks of the Ross Sea. Then the storyline simply stops. Skipping the denouement does not work cinematically — nor, given what happened next, does it do the film-makers' cause much good.

In essence, nothing happened. The 2012 CCAMLR meeting ended without a consensus, raising larger questions that the film might have addressed more directly: why do the limited economic and political benefits gained from exploitation of this fish continue to outweigh the seemingly innumerable ecological concerns? Is this marine ecosystem so remote that the preservation-minded are unlikely to see it and support its conservation? Some considered grappling with such issues would have been welcome.

A special CCAMLR meeting next month will revisit proposals for marine protected areas. Perhaps the added exposure from *The Last Ocean*, currently on the film-festival circuit, will turn negotiations around — and withhold the last laugh from Agent Smith. ■

Michael White *is* Nature*'s senior editor for climate science.*

Correspondence

Monster fern makes IUCN invader list

The list of 100 of the world's worst invasive alien species, compiled by the International Union for Conservation of Nature (IUCN), aims to help biodiversity conservation efforts worldwide (see go.nature.com/qa9z1g). After a position on the list fell vacant as a result of the global eradication of the rinderpest virus (see, for example, *Nature* **474,** 10–11; 2011), we coordinated the community of invasion biologists in a unique initiative to vote for a replacement.

We assessed more than 10,000 invasive species from the world's largest databases for their capacity to spread and for their potential ecological or economic impact. More than 650 experts from 63 countries then voted on the ten candidate species we shortlisted, and selected the giant salvinia (*Salvinia molesta*), an aquatic fern.

Native to Brazil, this fern has spread throughout the tropics and subtropics. It doubles in abundance within days, forming thick, floating mats that block light from expanses of water, reduce its oxygen content and degrade water quality. They also impede water-based transport, clog irrigation and power-generation systems, and harm local fisheries.

Now in the global spotlight, this new entrant to the IUCN list is set to increase public awareness of the harm caused by invasive species and to stimulate more discussion in science and policy circles.

Franck Courchamp* *CNRS; University of Paris-Sud, Orsay, France.*
franck.courchamp@u-psud.fr
**On behalf of 7 co-signatories. See go.nature.com/wvjef2 for full list.*

Avoid more organ transplant scandals

Our institution is launching an international transdisciplinary initiative to improve the lamentable state of solid-organ transplantation in Germany and to help fulfil society's obligations towards millions of organ donors and recipients worldwide (see go.nature.com/z5b7uo).

The mortality rate following liver transplantation has risen alarmingly across the country over the past few years. The survival rate after one year is only 72% in Germany, which is 20% lower than in the United States and the United Kingdom, even though Germany has more transplant centres and fewer organ donors per capita (see go.nature.com/pgmrpn; in German). Such scandals are leading to a steady decline in altruistic organ donations, with an 18% drop in the first quarter of this year compared with the same period in 2012.

The situation largely reflects the weak regulation of organ transplantation in Germany, especially by comparison with other countries such as the Netherlands and Denmark (C. Metz and N. Hoppe *Eur. J. Health Law* **20,** 113–116; 2013). Proposals to rectify this include setting up an independent institute of transplantation medicine that has regulative and standard-setting powers (see go.nature.com/yzl4km; in German).

A lack of good prognostic models compounds the likelihood of transplantation failure. Such models would allow clinical urgency to be weighed against transplantation outcome. We are therefore planning systematic multicentre trials to evaluate the prognostic value of liver-allocation scores.

Our initiative also aims to address the dearth of quality-management systems that are properly founded on comprehensive, evidence-based data and on precise methodology that considers patients' needs and expectations.

The early results are promising. We intend to publish regular updates to provide essential information to the transplantation community and to ensure openness to the public.

Harald Schrem *Hannover Medical School, Germany.*
schrem.harald@mh-hannover.de
Alexander Kaltenborn *Hannover Medical School; and Federal Armed Forces Medical Centre, Hannover, Germany.*

Satellites: make data freely accessible

The cost of accessing satellite data is hampering the widespread application of satellite monitoring, a vital tool for controlling deforestation (Jim Lynch *et al. Nature* **496,** 293–294; 2013) and for biodiversity assessments. We urge government agencies that produce taxpayer-funded satellite images to make these available free of charge and in user-friendly formats.

Lynch and colleagues' call for daily satellite observations of forests worldwide would mean aggregating information from numerous satellites that are operated by many countries. Assembling the large data sets needed for global monitoring would be prohibitively expensive, however, because national governments do not have a free-access policy for their satellite images.

One solution would be to combine data from the US Landsat satellites with those from the European Space Agency's planned Sentinel-2 satellites, which could deliver optical imagery with global coverage every 3–5 days. The distribution of Landsat imagery has increased by two orders of magnitude since 2008, when the US Geological Survey made all the data free to access online. Data from NASA's MODIS and all of their Earth-observation imagery are also available for free, as are data from the China–Brazil Earth Resources Satellite programme.

Woody Turner* *Earth Science Division, NASA, Washington DC, USA.*
woody.turner@nasa.gov
**On behalf of 14 co-signatories. See go.nature.com/pfv6an for full list.*

Satellites: ambition for forest initiative

We disagree strongly with the suggestion by Jim Lynch and colleagues that the outputs of the Global Observation of Forest and Land Cover Dynamics panel and the Global Forest Observations Initiative "lack ambition and an understanding of the potential of satellites" (*Nature* **496,** 293–294; 2013).

As participants in these programmes and in the United Nations Programme on Reducing Emissions from Deforestation and Forest Degradation (UN-REDD), we aim to show how remote sensing can help systematic global monitoring to make REDD+ a reality in the context of wider societal engagement. (REDD+ is a climate-mitigation initiative under the UN Framework Convention on Climate Change (UNFCCC).)

We also question the feasibility of Lynch and colleagues' call for rapid-response satellite monitoring of deforestation to be enshrined in international law under the UNFCCC, given national sovereignty concerns and the fact that we are not yet in a position to mitigate the problems of cloud cover. Although radar can penetrate cloud, the technology cannot yet capture changes in forest ecosystems in a systematic and repeatable way.

Giles Foody* *University of Nottingham, UK.*
giles.foody@nottingham.ac.uk
**On behalf of 7 co-signatories. See go.nature.com/wu1f3e for full list.*

CONTRIBUTIONS

Correspondence may be sent to **correspondence@nature.com** after consulting the author guidelines at **http://go.nature.com/cmchno**.

FORUM Theoretical physics

Sizing up atoms

Niels Bohr's model of the structure of the atom raised the question of how large an atom can be. One hundred years on, the issue is still unresolved. Two physicists discuss the theoretical limits of atomic and nuclear size.

Orbital arguments

PAUL INDELICATO

Bohr's model of the atom[1] (Fig. 1) provided a new way of thinking about atomic size. For example, it predicted that the radius of the smallest atom, hydrogen, in its ground state was 0.5×10^{-10} metres, 100,000 times larger than the size of the nucleus. This value, known as the Bohr radius (a_0), was remarkably accurate and is now one of the fundamental constants of atomic physics. The model also proposed that the speed of an electron in the inner orbital of an atom was approximately $Zc\alpha$ (where Z is the proton number, c is the speed of light and α is the fine-structure constant, approximately 1/137). Intriguingly, this limits Z to a maximum of about 137, because, above this value, the electron's speed would be greater than the speed of light.

Nowadays, atomic models are based on the Dirac equation, which combines relativity and quantum mechanics in a theory called quantum electrodynamics (QED). The Dirac equation for a point nucleus leads to the same limit: the electron-binding energy becomes complex when Z is greater than or equal to $1/\alpha$. But for an extended nucleus, the limit is around $Z = 173$. Above that value, the electron-binding energy is more than twice the electron's rest mass, a condition that allows the formation of electron–antielectron pairs, which would render the atom unstable.

The size of an atom can be defined in different ways[2]. If the mean spherical radius of the whole atom is considered, based on the total electron density, then the possible range of sizes is small: from $1.06a_0$ to $1.5a_0$. But if the size of the outermost orbital is considered, then atomic size ranges from a_0 at $Z = 1$ to $8a_0$ at $Z = 172$ (refs 2,3). What happens above $Z = 172$ is still being investigated[4] to study how the emission of real electron–antielectron pairs causes the breakdown of the quantum vacuum — a mysterious state predicted by QED, consisting of empty space in which virtual particles such as photons and electron–antielectron pairs are constantly created and annihilated.

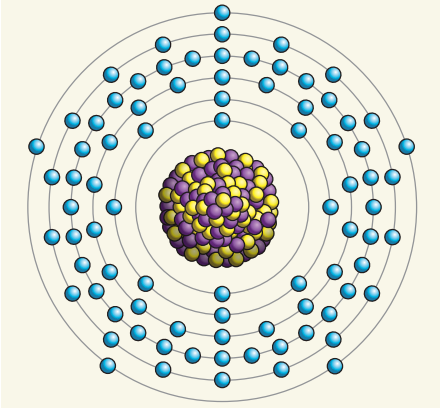


Figure 1 | Atomic structure. This cartoon of a bismuth-209 atom exemplifies the features of Niels Bohr's model of the atom: a nucleus, composed of protons (purple) and neutrons (yellow) is orbited by electrons (blue), which occupy distinct shells. The relative size of the nucleus and the electron shells are not shown to scale. Bismuth-209 can decay by emitting α-radiation, but the measured half-life[13] for this process is $1.9 \pm 0.2 \times 10^{19}$ years, a billion times longer than the estimated age of the Universe. It is, therefore, essentially the heaviest naturally occurring stable atom.

The heaviest nucleus to have been identified[5] has $Z = 118$. Nuclei with higher numbers of protons can be studied only by creating them temporarily during collisions of two lower-charged nuclei. This was attempted in the 1980s, but the accelerators of the time could not produce bare nuclei (or nuclei with single electrons) that had large enough Z to succeed. Today, high-quality beams of bare nuclei of heavy elements can be produced at energies that could allow binary nuclear systems to be prepared for approximately 10^{-21} seconds. Projects to study the quasi-molecular state created in such collisions, and to investigate the properties of the resulting quasi-atoms, have been proposed.

But it is not only large atoms that can be made — smaller, exotic atoms can also be created by replacing electrons with heavier particles, such as muons, pions or antiprotons. The resulting systems are 207 to 1,836 times smaller than the corresponding 'normal' atoms, and are thus close to the size of a nucleus. Such atoms have been used to study nuclear properties, such as the size of a proton[6].

Paul Indelicato *is at the Kastler Brossel Laboratory, ENS, CNRS, Université Pierre et Marie Curie, 75005 Paris, France.*
e-mail: paul.indelicato@lkb.upmc.fr

The nuclear question

ALEXANDER KARPOV

The maximum size of an atomic nucleus is determined by its stability towards decay. In general, only a few isotopes of each element are stable, the heaviest of which is bismuth-209 (83 protons and 126 neutrons; Fig. 1). All elements heavier than this are radioactive, although two of them (thorium and uranium) have tremendously long half-lives and are found in large amounts in nature. In some respects, such long-lived radioactive elements can be thought of as 'stable'.

If we enlarged a nucleus by adding neutrons, then it would become increasingly short-lived, eventually reaching the border of neutron stability. Beyond this border, the nuclear system is unbound and spontaneously emits neutrons. The number of neutrons that can be added to a stable nucleus depends mainly on Z: the larger Z is, the further away is the border of neutron stability. The border has already been reached experimentally in elements up to oxygen (and probably up to aluminium, although some dispute this). But for heavy elements, only theoretical estimates of the border's position are available. For example, the heaviest uranium atom is predicted to bind 92 protons and about 208 neutrons, a total mass number of around 300; by comparison, the heaviest naturally occurring uranium

nucleus has a mass number of 238.

If we added protons to uranium, the heaviest naturally occurring element, then we would produce new elements. (In fact, we would need to add protons and neutrons, to avoid reaching the border of proton stability). The resulting nuclei would be progressively less stable to spontaneous fission because of Coulomb repulsion in their interiors. Nuclei become totally unstable towards fission at about $Z = 106$, in the absence of quantum effects.

But nuclei consisting of certain 'magic' numbers of protons and neutrons are especially stable by comparison with their neighbours. Superheavy nuclei that have nearly magic numbers of protons and neutrons form islands of relatively long-lived nuclei surrounded by a sea of short-lived nuclei. A pair of magic numbers in the superheavy region (114 protons and 184 neutrons) was predicted[7–10] in the 1960s. The centre of this island has not been reached experimentally, and the ways to reach it are debated[11]. However, elements up to $Z = 118$ have been synthesized[5,12]. The existence of the island unambiguously follows from these results, but the data do not indicate where the top of the island is, nor how long-lived the nuclei at the top would be. No consensus on this topic has been reached from theoretical considerations.

Are there other islands of stability? Probably, yes. But different theories of nuclear stability diverge from each other when extrapolated into remote domains of nuclei, so the opposite answer cannot be excluded. One hypothesis proposes that very heavy nuclei do not have a 'normal', nearly uniform distribution of nuclear matter, but a bubble-like distribution. This should substantially suppress the Coulomb forces and increase nuclear stability. Some theories predict bubble-like structures in the vicinity of the first island of stability of superheavy nuclei — in which case, massive, long-lived nuclei might have rather exotic structures. ■

Alexander Karpov *is at the Flerov Laboratory of Nuclear Reactions, Joint Institute for Nuclear Research, 141980 Dubna, Moscow region, Russian Federation.*
e-mail: karpov@jinr.ru

1. Bohr, N. *Phil. Mag.* **26,** 1–25 (1913).
2. Indelicato, P., Santos, J. P., Boucard, S. & Desclaux, J.-P. *Eur. Phys. J. D* **45,** 155–170 (2007).
3. Indelicato, P., Bieroń, J. & Jönsson, P. *Theor. Chem. Acc.* **129,** 495–505 (2011).
4. Ackad, E. & Horbatsch, M. *Phys. Rev. A* **78,** 062711 (2008).
5. Oganessian, Y. T. *et al. Phys. Rev. C* **74,** 044602 (2006).
6. Pohl, R. *et al. Nature* **466,** 213–216 (2010).
7. Mosel, U., Fink, B. & Greiner, W. in *Memorandum zur Errichtung eines gemeinsamen Ausbildungszentrums fuer Kernphysik der Hessischen Hochschulen* (1966).
8. Mosel, U. & Greiner, W. *Z. Phys.* **222,** 261–282 (1969).
9. Meldner, H. *Ark. Fys.* **36,** 593 (1967).
10. Sobiczewski, A., Gareev, F. A. & Kalinkin, B. N. *Phys. Lett.* **22,** 500–502 (1966).
11. Zagrebaev, V. I., Karpov, A. V., Greiner, W. *J. Phys. Conf. Ser.* **420,** 012001 (2013).
12. Oganessian, Y. T. *et al. Phys. Rev. Lett.* **104,** 142502 (2010).
13. de Marcillac, P., Coron, N., Dambier, G., Leblanc, J. & Moalic, J.-P. *Nature* **422,** 876–878 (2003).

HIGH-TEMPERATURE SUPERCONDUCTIVITY

The sound of a hidden order

Ultrasound measurements in a copper oxide superconductor have revealed an exotic phase of matter, composed of loops of spontaneous quantum currents, that has hitherto excelled at evading observation. SEE LETTER P.75

JAN ZAANEN

Rigid things are obvious in the human world, but nature allows for circumstances in which hardness gets a quantum-physics twist. The electron systems formed in copper oxide compounds became famous with the discovery in 1986 that these materials become superconductors at high temperature. But this turned out to be only the tip of the iceberg: the intensive research that ensued revealed surprise after surprise. It became clear that the strongly interacting electrons of these systems form the building blocks of a plethora of exotic phases of matter that are shaped by the weirdness of quantum mechanics[1]. On page 75 of this issue, Shekhter *et al.*[2] present conclusive evidence for the existence of one such phase — one that breaks 'quantum-spookiness' records. Driven by a quantum effect known as zero-point motion, the electrons in this phase organize themselves into patterns formed from spontaneous current loops, and the phase transition in which this electronic order sets in leaves an unambiguous mark on the sound waves travelling through the copper oxide lattice.

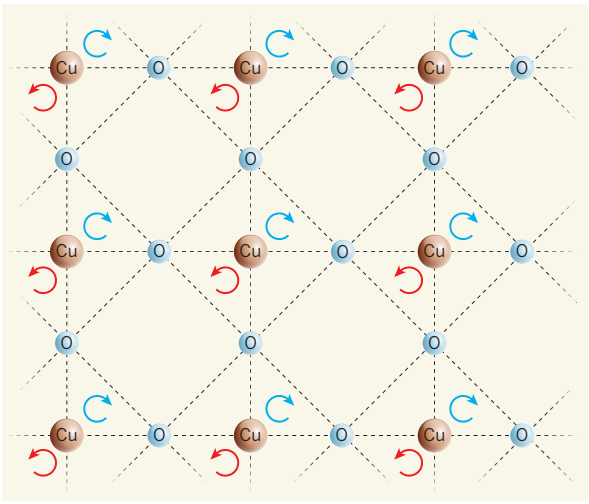


Figure 1 | Electronic order. Shekhter *et al.*[2] demonstrate an electronic order in a copper oxide compound (Cu, copper; O, oxygen) which consists of countercirculating currents (arrows) within the unit cells of the compound's atomic lattice.

The discovery of a phase of matter formed from spontaneous quantum currents is stunning in itself: this 'hidden order' has been playing hide-and-seek for a long time[1]. The first indications of it came from neutron-scattering experiments[3,4]. However, to qualify as a phase of matter, such an electronic order must set in suddenly at a critical temperature. Measuring thermodynamic quantities such as the specific heat is the standard way to detect such phase transitions, because at the critical temperature these quantities should show singularities — sharp cusps in their temperature dependence. These singularities have not been detected, but it was argued[4] that, given its special symmetry, this order could conceal itself completely even in this regard.

Much like the vibrating strings of a violin produce sound waves, the vibrations of the ions in copper oxide compounds also generate sound waves. At high (ultrasound) frequencies, such 'phonons' lose their energy to the electron system, and when the electrons undergo a phase transition, their 'boiling' markedly increases their capacity to damp the phonons. This is precisely what Shekhter *et al.* observe in their ultrasound measurements of the copper oxide compound $YBa_2Cu_3O_{6+\delta}$: at the critical temperature, the onset of the current-loop order in this material causes sharp changes in both the speed and the lifetime of the phonons. These changes reveal the thermodynamic singularities demonstrating that the currents form a macroscopic phase of matter.

What is the origin of this form of spontaneous-current order? Although details remain to be settled, theorists

have played a key part in guiding experimentalists to look in the right places, indicating that the underlying theory is trustworthy at least to a certain degree. The physics behind the current loops is counterintuitive, arising in the brew of quantum mechanics and strong interactions[1]. The chemistry of the copper oxides causes the electrons to repel each other so strongly, while their density is high, that they impede each other's motion. An appropriate metaphor is to view these systems as traffic jams, with the difference being that an electron's urge to move comes from the demand of eternal quantum motion.

Resting on the mathematical theory describing such quantized traffic jams, the idea was born[5] in the 1980s that the electrons might organize into a state with spontaneous currents, and in 1997 it was proposed[6] that the currents might form a pattern of counter-circulating flows inside the unit cell of the copper oxide lattice (Fig. 1) — which now seems to be confirmed by Shekhter and colleagues' measurements. This particular pattern of currents was inspired[6] by the state's capacity to hide, because the only symmetry it breaks is the eerie reversal of time[7,8].

This current order is sturdy: at low levels of hole (the absence of an electron) doping, at which the electronic traffic jam effects are particularly strong, the order sets in at quite high temperatures, whereas it gradually weakens when the doping increases, and disappears when superconductivity is strongest[1,4]. Could this order be the cause of superconductivity? It can be argued that the severe quantum fluctuations that develop when the current order disappears altogether as a function of doping might 'glue' electrons into Cooper pairs[4] — a key ingredient in superconductivity. But a lot goes on in the copper oxides besides loop currents and superconductivity[1]. The idea of exotic orders started in the 1990s with the observation of electronic stripes, a form of spatial self-organization of the electronic traffic jam[9], and since then claims of several other exotic ordering phenomena have been made[1].

The simultaneous presence of all of these different ordering tendencies in the copper oxides is not at all understood, and the mystery deepens further when the electrons are heated to temperatures well above the critical temperatures of the current order and of superconductivity. Here, all this complexity disappears, and instead a 'strange metal' phase is observed experimentally, which completely confounds the present understanding of quantum many-body theory[1]. Recent attempts to unleash the mathematics of string theory, in which particles are described by extended entities called strings, seem to shed light on this mystery. These 'AdS/CFT' calculations predict strange metals that are quite like those seen in the laboratory: at low temperatures, they turn into several competing orders, including superconductivity[10].

It might be that much will also be learned in this regard from Shekhter and colleagues' ultrasound measurements. Sound-wave propagation is affected by fluctuations in the electron system not only at the phase transitions that occur in these materials but over the whole range of doping and temperature in which the competing orders and the strange-metal phase occur. These data indicate that this electronic stuff is, in this whole regime, fluctuating under the influence of heat in a way that is utterly different from boiling matter in our everyday world. It may be that further analysis of these ultrasound data may unlock some of the deepest secrets of this mysterious 'quantum matter'. ■

Jan Zaanen *is at the Instituut Lorentz for Theoretical Physics, Leiden University, 2300 RA Leiden, the Netherlands.*
e-mail: jan@lorentz.leidenuniv.nl

1. Zaanen, J. in *100 Years of Superconductivity* (eds Rogalla, H. & Kes, P. H.) Ch. 2.4, 92–117 (Chapman & Hall, 2011); preprint at http:/arXiv.org/abs/1012.5461.
2. Shekhter, A. *et al. Nature* **498,** 75–77 (2013).
3. Fauqué, B. *et al. Phys. Rev. Lett.* **96,** 197001 (2006).
4. Varma, C. M. *Nature* **468,** 184–185 (2010).
5. Affleck, I. & Marston, J. B. *Phys. Rev. B* **37,** 3774 (1988).
6. Varma, C. M. *Phys. Rev. B* **55,** 14554–14580 (1997).
7. Kaminski, A. *et al. Nature* **416,** 610–613 (2002).
8. Xia, J. *et al. Phys. Rev. Lett.* **100,** 127002 (2008).
9. Zaanen, J. *Nature* **440,** 1118–1119 (2006).
10. Liu, H. *Phys. Today* **65**(6), 68–69 (2012).

IMMUNOLOGY

An innate regulatory cell

The finding that innate lymphoid cells can control the activity of CD4$^+$ T cells reveals another potential form of immune-system regulation, and may help to explain how the body distinguishes resident from pathogenic bacteria. SEE LETTER P.113

MARCO COLONNA

The lining of our intestines is a border zone at which our own cells peacefully coexist with resident bacteria. Cells of the immune system patrol this area to prevent infiltration by invasive pathogenic bacteria. However, in some individuals, the immune system mistakenly targets the benign commensal bacteria, triggering an inflammatory process that damages the intestinal mucosa and leads to inflammatory bowel disease[1,2]. In a report on page 113 of this issue, Hepworth *et al.*[3] identify a mechanism that could be crucial for preventing an overly exuberant immune response to commensal bacteria. Intriguingly, this mechanism relies on the ability of a rare type of immune cell — innate lymphoid cells — to control T cells of the adaptive immune system.

Innate lymphoid cells (ILCs) are classified into three groups, depending on their expression of and developmental dependence on certain transcription factors and secreted molecules. Hepworth and colleagues studied group 3 ILCs, which reinforce the intestinal barrier against pathogenic bacteria by producing the soluble molecules IL-22 and IL-17A. These cytokines augment the capacity of epithelial cells to produce antimicrobial peptides and also recruit other immune cells, such as granulocytes[4,5].

The development of group 3 ILCs is driven by the transcription factor RORγt (ref. 6), and so these cells are absent from RORγt-deficient mice. Hepworth *et al.* noted that RORγt-deficient mice had symptoms that were characteristic of immune activation: their spleens were enlarged and contained activated T cells expressing the CD4 receptor (CD4$^+$ T cells). Moreover, the serum of the mice contained antibodies that bind to commensal bacteria, suggesting specific immune activity against these bacteria. Consistent with this, the authors could ameliorate the CD4$^+$ T-cell activation by using antibiotic treatment to eliminate the commensal bacteria. The researchers saw a similar effect when they depleted the group 3 ILCs in normal mice, confirming that these cells are essential for controlling CD4$^+$ T-cell activation.

But how does this regulation occur? Although group 3 ILCs produce IL-22 and IL-17A, mice that lacked these cytokines did not have symptoms of immune activation. To gain more insight, Hepworth and colleagues turned to analysis of the transcriptome — a cell's complement of RNA molecules. This revealed that the genes that encode MHC class II proteins are highly expressed by a subset of group 3 ILCs called lymphoid tissue inducer (LTi)-like cells. LTi-like cells appear after birth in the intestinal mucosa, in which they promote the generation of post-natal lymphoid tissue by recruiting B cells to form isolated lymphoid follicles[7]. MHC class II molecules capture

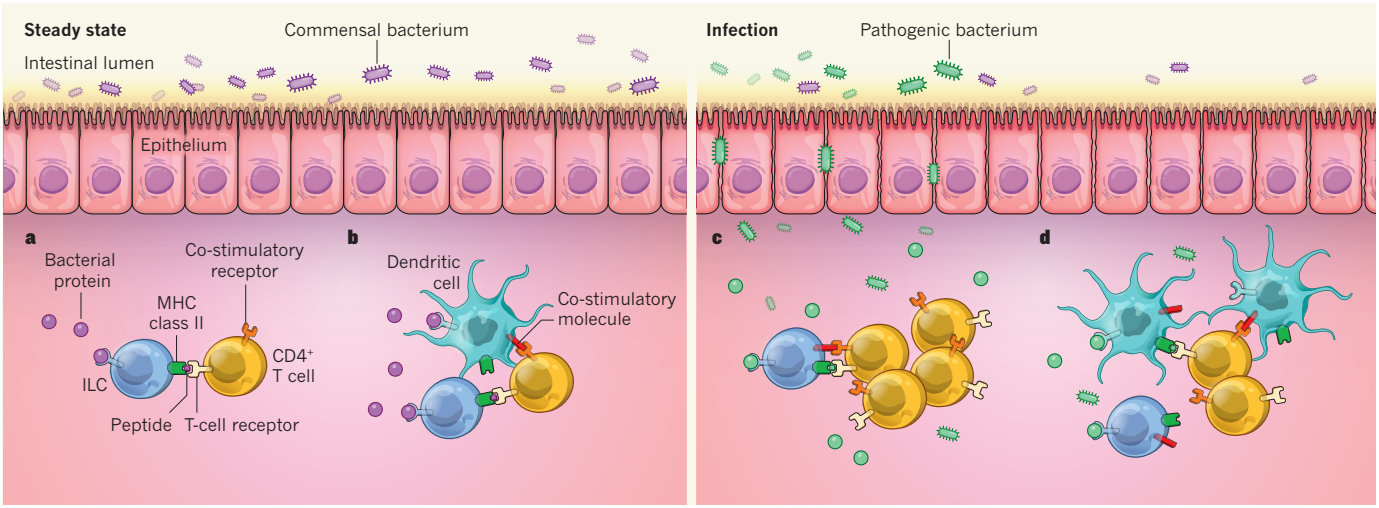


Figure 1 | Innate control of CD4⁺ T cells. Hepworth *et al.*[3] show that, during the steady state, group 3 innate lymphoid cells (ILCs) help to prevent an immune response from CD4⁺ T cells that express T-cell receptors specific for peptides derived from resident commensal bacteria in the intestine. **a**, The authors propose that ILCs interact with CD4⁺ T cells through the presentation of peptides on the MHC class II protein complex, and that this occurs in the absence of co-stimulation. **b**, Alternatively, ILCs might prevent dendritic-cell-induced CD4⁺ T-cell activation through competition. **c**, **d**, By contrast, ILCs do not inhibit CD4⁺ T-cell responses against invading pathogenic bacteria. This changed response might be explained by enhanced expression of co-stimulatory molecules on the ILCs (**c**), which are induced by cytokines produced by other immune cells in response to infections, or by the presence of more dendritic cells (**d**), which outcompete the ILCs.

protein fragments called peptides and present these antigens to CD4⁺ T cells, which then scrutinize the peptide–MHC complexes for the presence of 'self' or foreign (typically microbial) peptides. Thus, it seems possible that group 3 ILCs influence CD4⁺ T-cell activation through this antigen-presentation process.

Capture, processing and presentation of antigens is mainly a function of dendritic cells — 'sentinel' immune cells that instruct T cells to tolerate MHC molecules in complex with self peptides, but to react against foreign-peptide–MHC complexes[8]. Whether dendritic cells induce T-cell activation or tolerance depends on whether the dendritic cells are concurrently activated by microbial stimuli that induce the accessory expression of co-stimulatory molecules. Hepworth *et al.* demonstrate that group 3 ILCs can capture model proteins (chicken ovalbumin or Eα protein), degrade them, and present the peptides on MHC class II, just like dendritic cells. The group 3 ILCs could also present a peptide derived from commensal bacteria. However, the cells did not induce proliferation of CD4⁺ T cells that expressed receptors specific for these antigens, probably because presentation by MHC class II on ILCs is not coupled with co-stimulatory molecules and hence induces T-cell tolerance rather than activation.

To conclusively demonstrate the relevance of MHC class II expression by group 3 ILCs to the control of CD4⁺ T cells, Hepworth and colleagues generated mice that lacked MHC class II proteins in group 3 ILCs only. They observed signs of CD4⁺ T-cell activity against commensal bacteria, just as in RORγt-deficient mice. In addition, over time these mice spontaneously developed rectal prolapse, which is characteristic of inflammatory bowel disease.

Overall, Hepworth and colleagues' study reveals that group 3 ILCs can capture and present antigens much like dendritic cells, but that they regulate CD4⁺ T-cell function in a manner somewhat resembling that of another class of T cell, regulatory T cells. These findings are provocative and raise important questions. Where and how do group 3 ILCs capture microbial peptides or proteins? And where do the ILCs and CD4⁺ T cells interact? Both processes could take place in the mesenteric lymph node, in which bacterial products are collected from lymph and CD4⁺ T cells are initially activated. Alternatively, group 3 ILCs might act in the intestinal mucosa, where they are in close contact with commensal bacteria and where CD4⁺ T cells perform most of their effector functions.

Details on the regulatory function of ILCs also remain uncertain. Do these cells induce CD4⁺ T-cell death or functional paralysis, or do they actively suppress CD4⁺ T cells, like regulatory T cells do? Hepworth *et al.* propose that group 3 ILCs interact directly with CD4⁺ T cells (Fig. 1a). However, it is possible that the ILCs compete with dendritic cells for interaction with CD4⁺ T cells, and thereby prevent their activation (Fig. 1b). Group 3 ILCs might also enable regulatory interactions of CD4⁺ T cells with other cells by ensuring an appropriate architecture of the lymphoid tissue in the intestinal mucosa.

Finally, how do the group 3 ILCs inhibit CD4⁺ T-cell responses to commensal bacteria but not to pathogenic bacteria? ILCs have been shown to promote the memory function of CD4⁺ T cells through the co-stimulatory molecules OX40L and CD30L[9]. Perhaps pathogenic infection and the consequent release of inflammatory cytokines (such as IL-23) enhance expression of these co-stimulatory molecules on ILCs, thereby changing their interactions with CD4⁺ T cells (Fig. 1c). In this regard, it would be interesting to know the status of ILCs during inflammation. Alternatively, recruitment and activation of dendritic cells might overwhelm the regulatory ILCs during infection (Fig. 1d). Further delineation of the regulatory and immunogenic functions of innate lymphoid cells, including group 3 ILCs, will not only help us to understand immune regulatory processes but could also provide the basis for new, refined therapeutic intervention in conditions such as inflammatory bowel disease. ■

Marco Colonna *is in the Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA.*
e-mail: mcolonna@pathology.wustl.edu

1. Hooper, L. V., Littman, D. R. & Macpherson, A. J. *Science* **336,** 1268–1273 (2012).
2. Maloy, K. J. & Powrie, F. *Nature* **474,** 298–306 (2011).
3. Hepworth, M. R. *et al. Nature* **498,** 113–117 (2013).
4. Pappu, R., Rutz, S. & Ouyang, W. *Trends Immunol.* **33,** 343–349 (2012).
5. Ouyang, W., Rutz, S., Crellin, N. K., Valdez, P. A. & Hymowitz, S. G. *Annu. Rev. Immunol.* **29,** 71–109 (2011).
6. Eberl, G. & Littman, D. R. *Immunol. Rev.* **195,** 81–90 (2003).
7. Ivanov, I. I., Diehl, G. E. & Littman, D. R. *Curr. Top. Microbiol. Immunol.* **308,** 59–82 (2006).
8. Steinman, R. M., Hawiger, D. & Nussenzweig, M. C. *Annu. Rev. Immunol.* **21,** 685–711 (2003).
9. Withers, D. R. *et al. Immunol. Rev.* **244,** 134–148 (2011).

TECHNIQUES

Optical spectroscopy goes intramolecular

Optical spectroscopic imaging has taken a leap into the intramolecular regime with an approach that achieves subnanometre spatial resolution. The technique should find applications in photochemistry and nanotechnology. SEE LETTER P.82

JOANNA M. ATKIN & MARKUS B. RASCHKE

The goal of optical microscopy is to visualize the physical and chemical properties of objects too small to be seen with the naked eye. However, objects separated by less than approximately half the wavelength of the light that is used to illuminate them can in general not be distinguished, owing to the inherent wave nature of light. The development of near-field optics has broken this spatial-resolution limit and has enabled optical imaging and spectroscopy with a resolution of a few nanometres. On page 82 of this issue, Zhang *et al.*[1] report an optical spectroscopic imaging approach that achieves subnanometre resolution and resolves the internal structure of a single molecule.

In 1928, Edward Hutchinson Synge came up with an idea for nanometre-scale optical microscopy[2], possibly inspired by Richard Zsigmondy's ultramicroscope[3]. Synge suggested that light scattered by a small particle placed close to an object could act as a localized light source. Spatial resolution would then be determined by the size of the particle rather than the wavelength of the light used.

The experimental implementation of this idea, however, had to await the invention of scanning tunnelling microscopy (STM) in the 1980s, because of the need for precise, nanometre-scale spatial control of the sample and scatterer. STM, which is based on a quantum tunnelling current of electrons between a nanoscale tip and the sample, provided spatial resolution down to the atomic scale[4]. This breakthrough was followed by the development of atomic force microscopy (AFM). Because AFM does not rely on a tunnelling current, it can be used on a much wider range of samples, including non-conducting materials and soft matter.

Although STM, AFM and other techniques such as transmission electron microscopy and X-ray microscopy can achieve results with atomic resolution, the goal of reaching this ultrahigh resolution, in combination with the detailed and sensitive information that optical spectroscopy would provide, remained unachieved. The power of optical spectroscopy lies in its sensitivity to energetic details of the configuration and electronic structure of atoms and molecules in solids, and the way in which these fundamental properties are coupled. In particular, when probing vibrational motion between atoms, optical spectroscopy can be used to identify the chemical constituents of molecules and solids.

Possible ways of combining STM and AFM with optical techniques to provide nanometre-scale spectroscopic information have been extensively explored. The initial approach of near-field scanning optical microscopy (NSOM), which is based on the use of a tapered fibre with an STM or AFM feedback mechanism for controlling the sample–tip distance on the nanometre scale, provided[5] spatial resolution to below 100 nm. Other methods that paved the way to higher resolution and greater versatility included scanning plasmon near-field microscopy[6] and photonic force microscopy[7]. These developments led to the technique of scattering scanning near-field optical microscopy (*s*-SNOM). This generalization of NSOM and the early methods provides the most versatile realization of Synge's vision[8]. In *s*-SNOM, the apex of the tip (preferably metallic) serves as the nanoscale scatterer, enabling almost any optical spectroscopy technique to be extended to near-field use for probing electronic and vibrational properties with a spatial resolution of 10 nm or better[9,10]. (Following the development of near-field microscopy, powerful, super-resolution far-field optical microscopic techniques emerged, but these have typically provided limited spectroscopic information.)

Meanwhile, STM has been extended to yield vibrational and thus chemical spectroscopic information with atomic resolution using an approach called inelastic tunnelling spectroscopy[11]. Although so far limited to operating at cryogenic temperature conditions, this technique set the stage for what is possible in terms of spatial resolution and spectral content.

Zhang *et al.* extend these previous efforts by combining low-temperature STM (78 kelvin) in an ultrahigh vacuum with Raman spectroscopy as an optical vibrational spectroscopy technique (Fig. 1). In Raman spectroscopy, incident laser photons lose energy to specific molecular vibrational excitations in the sample, thus providing chemical 'fingerprints'. The combination of Raman spectroscopy with specially designed silver or gold STM tips, which can confine and locally enhance the incident laser field at the apex, is called tip-enhanced Raman scattering (TERS). Using STM and silver tips, Zhang and colleagues achieved subnanometre spatial resolution and were able to map spectroscopic signatures inside a single molecule, and to determine how these signatures changed with molecular orientation.

Optical spectroscopy with atomic-scale spatial resolution previously seemed impossible, with *s*-SNOM and TERS thought to be limited by the depth to which light can penetrate into the metallic tip — on the order of 10 nm at visible and infrared wavelengths. However, optical fields can be confined to almost arbitrarily small regions[12], which are limited only by the size at which the electrons in a homogeneous medium cease to

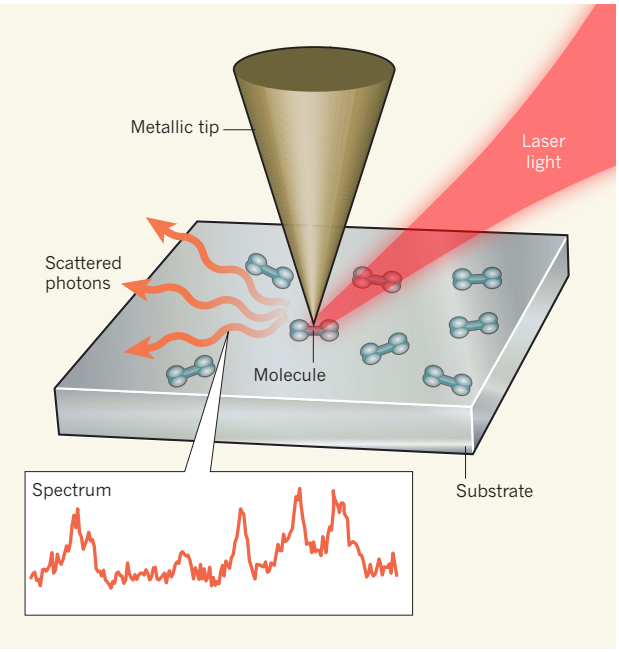


Figure 1 | Optical spectroscopic nano-imaging. Zhang and colleagues[1] have resolved the internal structure of a single molecule on the surface of a substrate by optical spectroscopy. They hold a metallic tip that has a very sharp apex (a few nanometres across) in close proximity to the molecule, and monitor the electric current (not shown) arising from electrons tunnelling between the tip and the sample. Laser light is then focused on the apex. Detection of the resulting tip-scattered photons provides a vibrational spectroscopic signature of the molecular structure.

behave as free particles. This limit is given by the Thomas–Fermi screening length of about 0.1 nm, below which non-local effects become significant. The fact that scanning near-field optical microscopy techniques have not previously achieved such high spatial resolution is probably due to the AFM and STM instruments used for optical techniques not having been designed with atomic resolution in mind.

The mechanistic details underlying the unprecedented optical resolution and molecular sensitivity obtained in Zhang and colleagues' work is not yet completely clear. The TERS signal measured seems to increase nonlinearly with increasing power of the incident laser, in contrast to what is observed with conventional Raman or TERS spectroscopy. The authors attribute this to a higher-order nonlinear response generating the signal. Moreover, the TERS signal was found to be sensitive to the optical properties of the tip in unexpected ways. The combination of these factors raises questions for theory and calls for further investigation.

The authors' work opens up avenues for probing and even controlling materials on molecular scales. Because it can be combined with essentially any optical technique, detailed specific chemical and physical information about many kinds of samples can be obtained, with the only limitation being the requirement of STM for an electrically conducting sample. The highly localized laser-field enhancement can also be used for photochemistry on the nanoscale, making and breaking bonds on the molecular level. Ultimately, this development could lead to new techniques for probing and controlling nanoscale structure, dynamics, mechanics and chemistry. ■

Joanna M. Atkin *and* **Markus B. Raschke**
are in the Department of Physics, Department of Chemistry, and JILA, University of Colorado at Boulder, Boulder, Colorado 80309-0390, USA.
e-mail: markus.raschke@colorado.edu

1. Zhang, R. *et al. Nature* **498,** 82–86 (2013).
2. Synge, E. H. *Phil. Mag.* **6,** 356–362 (1928).
3. Siedentopf, H. & Zsigmondy, R. *Ann. Phys.* **315,** 1–39 (1902).
4. Binnig, G., Rohrer, H., Gerber, C. & Weibel, E. *Phys. Rev. Lett.* **50,** 120–123 (1983).
5. Pohl, D. W., Denk, W. & Lanz, M. *Appl. Phys. Lett.* **44,** 651–653 (1984).
6. Specht, M., Pedarnig, J. D., Heckl, W. M. & Hänsch, T. W. *Phys. Rev. Lett.* **68,** 476–479 (1992).
7. Florin, E.-L., Pralle, A., Hörber, J. K. H. & Stelzer, E. H. K. *J. Struct. Biol.* **119,** 202–211 (1997).
8. Atkin, J. M., Berweger, S., Jones, A. C. & Raschke, M. B. *Adv. Phys.* **61,** 745–842 (2012).
9. Cialla, D. *et al. J. Raman Spectrosc.* **40,** 240–243 (2009).
10. Yano, T., Verma, P., Saito, Y., Ichimura, T. & Kawata, S. *Nature Photon.* **3,** 473–477 (2009).
11. Stipe, B. C., Rezaei, M. A. & Ho, W. *Science* **280,** 1732–1735 (1998).
12. Kreibig, U. & Vollmer, M. *Optical Properties of Metal Clusters* (Springer, 1995).

BIOCHEMISTRY

The ylide has landed

The enzyme co-substrate SAM has long been known to have two chemically distinct roles. A study of the CmoA enzyme suggests that SAM has a third trick up its sleeve — it forms species known as ylides. SEE LETTER P.123

BRADLEY J. LANDGRAF & SQUIRE J. BOOKER

Rarely has nature made such efficient use of a compound as it has of the biomolecule *S*-adenosylmethionine (SAM). SAM contains a positively charged sulphur atom known as a sulphonium group, which means that this molecule is often used as an electrophile — a polar species that is attracted to electron-rich centres. The compound also initiates a host of non-polar biochemical transformations that are mediated by free radicals[1,2]. But sulphonium groups have another ability that so far has not been observed in biochemical transformations: they can promote reactions by forming dipolar 'ylide' intermediates, which act as nucleophiles by reacting with electron-poor centres. In this issue, Kim *et al.*[3] (page 123) report strong evidence that an ylide intermediate is formed from SAM in the biosynthesis of a modified nucleotide, 5-oxyacetyl uridine*.

Ylides contain two opposing charges on adjacent atoms. In most ylides, a carbon atom containing an unshared pair of electrons is bonded to a positively charged atom, usually nitrogen, phosphorus or sulphur. Sulphonium-containing ylides are routinely used in synthetic organic chemistry, particularly to prepare molecules that contain small rings of atoms. Although they have been proposed as intermediates in several biochemical transformations[4–7], there has been no compelling evidence for this role.

5-Oxyacetyl uridine (cmo5U) is formed by the post-transcriptional modification of uridines (RNA bases) that occupy 'wobble' positions in several bacterial transfer RNAs.

*This article and the paper under discussion[3] were published online on 15 May 2013.

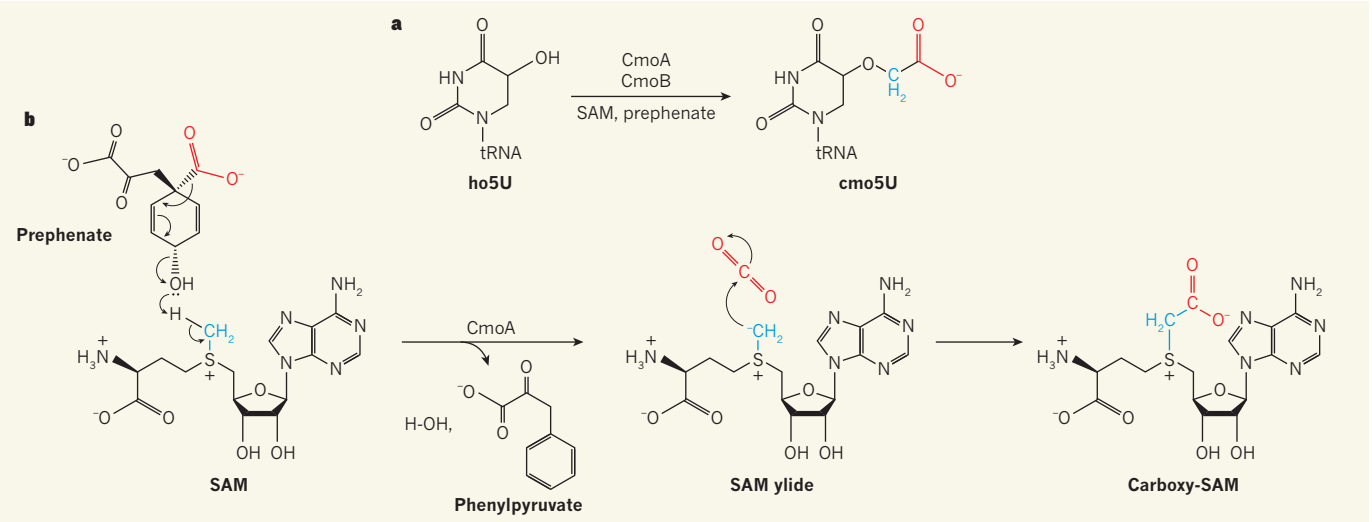


Figure 1 | An ylide biosynthetic intermediate. **a**, In the biosynthesis of the post-transcriptional modification 5-oxyacetyl uridine (cmo5U), the enzymes CmoA and CmoB catalyse the addition of an acetate group (blue and red) to the hydroxyl group (OH) of 5-hydroxy uridine (ho5U). C2 of the acetate (blue) comes from the co-substrate *S*-adenosylmethionine (SAM). Kim *et al.*[3] propose that C1 (red) comes from another co-substrate, prephenate (tRNA, transfer RNA). **b**, The authors suggest that prephenate loses its carboxylate group (CO_2^-) as a molecule of carbon dioxide in a process that results in the formation of a SAM ylide (a species in which positive and negative charges reside on adjacent atoms). Phenylpyruvate is generated as a side product. The ylide reacts with the CO_2 to form carboxy-SAM, in which an acetate is attached to the sulphur atom of SAM. The acetate is then transferred to ho5U in the presence of CmoB to form cmo5U (reaction not shown). Curved arrows indicate electron movement.

This modification allows a single transfer RNA to decode all the possible codons (three-nucleotide sequences) that could encode a particular amino acid. Genes that encode the SAM-dependent enzymes CmoA and CmoB are required for the biosynthesis of cmo5U, and 5-hydroxy uridine (ho5U) is probably an intermediate in the pathway[8] (Fig. 1a). Inactivation of the *cmoA* gene results in the formation of the incompletely modified tRNA bases ho5U and methoxy uridine, whereas only ho5U is formed on inactivation of *cmoB*. Perplexingly, genetic and biochemical studies[9] indicate a role in the generation of cmo5U for a metabolite that originates from the chorismate biosynthetic pathway, which is key to the formation of certain amino acids and for other crucial metabolites.

Formally, the biosynthesis of cmo5U from ho5U involves the attachment of the second carbon atom (C2, in the methyl group) of an acetate unit to the hydroxyl group (OH) of ho5U, rather than the first carbon atom (C1, in the carboxylate group; Fig. 1a). But these groups are unlikely to react with each other because they both tend to form nucleophilic, negatively charged species that are chemically incompatible. Moreover, isotopic labelling studies[9] have shown that the C2 carbon of the acetate moiety in cmo5U derives from the methyl group of SAM, which, in turn, derives from the amino acid methionine. This begs the question: what is the origin of the C1 carbon?

Kim *et al.* have solved the X-ray crystal structure of CmoA from the bacterium *Escherichia coli* and found an unexpected treasure buried in the enzyme's active site: a carboxy-SAM molecule, in which a carboxylate group (CO_2^-) is covalently attached to the methyl group of SAM. So how did the carboxylate group become attached?

The sulphur atom of SAM is bonded to carbon atoms in three chemical entities: a methyl group, a 5′-deoxyadenosyl group and a 3-amino-3-carboxypropyl group (Fig. 1b). The positive charge also causes the hydrogens on these adjacent carbon atoms to be weakly acidic — they can be removed as protons (H^+ ions) to form ylides. Carboxy-SAM could therefore be generated if the ylide that forms by deprotonation of SAM's methyl group attacks some electrophilic source of a carboxylate group.

However, the crystal structure described by Kim and colleagues reveals no group in the carboxy-SAM-binding pocket that could deprotonate the methyl group of SAM. Moreover, when the authors tested common electrophilic sources of carboxylates in the *in vitro* reaction of CmoA, none was effective. The chorismate biosynthetic pathway is known to be important in the formation of cmo5U; on this basis, the authors tested chorismate (an amino-acid precursor) as a potential carboxylate donor. Sure enough, they observed the slow formation of carboxy-SAM.

The researchers also observed another product, phenylpyruvate. They hypothesized that this product formed in two steps: an uncatalysed rearrangement of chorismate, which forms a compound called prephenate; then a CmoA-catalysed decarboxylation reaction in which prephenate releases a molecule of carbon dioxide (Fig. 1b). When the authors tested prephenate as a carboxylate donor, the CmoA reaction was much faster and higher yielding than it was with chorismate, and proceeded without the lag period observed with chorismate. Moreover, when chorismate labelled with carbon-13 was used in the reaction, the carbon label transferred to the carboxylate of carboxy-SAM.

On the basis of the above observations, Kim *et al.* propose that prephenate loses CO_2 and eliminates a hydroxide ion (OH^-), which is sufficiently basic to remove a proton from the methyl group of SAM, generating a nucleophilic ylide (Fig. 1b). The ylide then reacts with the liberated CO_2 to give carboxy-SAM, in which the C2 carbon of the acetate moiety — the same carbon that was nucleophilic in the ylide — is electrophilic. The authors went on to perform studies with radiolabelled SAM, providing evidence to support a reversible ylide formation that requires hydroxide ions.

The researchers then conducted the CmoA reaction in the presence of CmoB and tRNA purified from a *cmoB*-deficient strain of *E. coli* — that is, ho5U-containing RNA — and observed the formation of cmo5U. The generation of this product involves the attack of the nucleophilic hydroxyl group of ho5U on carboxy-SAM. Overall, it seems that nature cleverly uses the sulphonium moiety of SAM to invert the reactivity of the C2 of an acetate unit from nucleophilic to electrophilic, so that a suitable nucleophile can attack it.

The decarboxylation of prephenate to generate both a strong nucleophile (OH^-) and a strong electrophile (CO_2) could cause disastrous side reactions. The way in which the active site carefully orchestrates its catalytic sequence will therefore be of utmost interest to biologists. Kim and co-workers used computational modelling to dock prephenate in the active site of CmoA, and found that the methyl group of SAM is sandwiched between the hydroxyl and carboxylate groups

of prephenate. However, the hydroxyl group is poorly positioned for nucleophilic attack on the methyl group, although it is appropriately positioned for proton removal. Further X-ray structures of CmoA in complex with prephenate and/or unreactive SAM analogues might shed light on the details of crucial subtleties in the reaction.

The other interesting aspect of the cmo5U post-translational modification is the mechanism by which a hydroxyl group is attached to uridine to form ho5U, especially in the absence of molecular oxygen. An electrophilic source of the hydroxyl group would be required for this reaction — it would be amazing if this group were supplied by another unexpected metabolite. ■

Bradley J. Landgraf *and* **Squire J. Booker**
are in the Departments of Chemistry and of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.
e-mail: squire@psu.edu

1. Challand, M. R., Driesener, R. C. & Roach, P. L. *Nat. Prod. Rep.* **28,** 1696–1721 (2011).
2. Booker, S. J. *Curr. Opin. Chem. Biol.* **13,** 58–73 (2009).
3. Kim, J. *et al. Nature* **498,** 123–126 (2013).
4. Iwig, D. F., Grippe, A. T., McIntyre, T. A. & Booker, S. J. *Biochemistry* **43,** 13510–13524 (2004).
5. Iwig, D. F. & Booker, S. J. *Biochemistry* **43,** 13496–13509 (2004).
6. Kinzie, S. D., Thern, B. & Iwata-Reuyl, D. *Org. Lett.* **2,** 1307–1310 (2000).
7. Tokiwano, T., Watanabe, H., Seo, T. & Oikawa, H. *Chem. Commun.* 6016–6018 (2008).
8. Nasvall, S. J., Chen, P. & Björk, G. R. *RNA* **13,** 2151–2164 (2004).
9. Hagervall, T. G., Jonsson, Y. H., Edmonds, C. G., McCloskey, J. A. & Björk, G. R. *J. Bacteriol.* **172,** 252–259 (1990).

CLIMATE SCIENCE

Plant a tree, but tend it well

Forests recovering from human disturbance act as a substantial sink that helps to absorb anthropogenic carbon dioxide emissions. Simulations suggest that nutrient limitation reduces that effect.

JULIA PONGRATZ

An understanding of the global carbon cycle lies at the core of our understanding of climate change. The atmospheric concentration of carbon dioxide strongly influences global temperatures, and changes in this concentration result from variations in carbon sinks (in the ocean and the land biosphere) and sources (fossil-fuel burning and the transformation of natural ecosystems, such as forests, to managed land). The largest uncertainty in all of these carbon fluxes is associated with anthropogenic emissions that arise from changes in land use. Writing in *Global Change Biology*, Jain *et al.*[1] report that land-use emissions might have been substantially underestimated because the effects of nutrient limitation on plant growth were disregarded.

Any gardener will tell you that a lack of nutrients, such as nitrogen, limits the growth of plants — it is easy to plant a tree, but much harder to make it grow. The problem is exacerbated by harvesting, which removes nutrients from the soil. Nutrients can be replaced by soil microbes as they decompose organic material, by plants that fix nitrogen from the air, by the addition of fertilizers or by the deposition from the atmosphere of nitrogen-containing compounds produced by industrial processes (one of the few beneficial side effects of air pollution). The rate of nutrient supply limits plant regrowth. The evidence is not unequivocal, but as increases in the atmospheric concentration of CO_2 stimulate plants to grow faster, ecosystems could become limited by nutrient availability, thereby reducing their potential as carbon sinks[2].

Observed land-use emissions represent the carbon released when a forest is cleared minus the CO_2 taken up by vegetation as it regrows. Today, forest is regrowing on abandoned agricultural land in Eurasia and North America, and worldwide in areas where wood has been harvested. Jain *et al.* used a numerical model of global terrestrial vegetation to test how strongly vegetation regrowth over the past century has been impeded by nitrogen limitation. The authors found that when nitrogen limitation is accounted for in their model, the uptake of CO_2 by regrowing vegetation slows, and that simulated global land-use emissions are 40% higher than in a model that disregards nitrogen limitation (Fig. 1). The effect is small in the tropics, where the moist, warm climate stimulates microbes to liberate nutrients from leaf litter quickly, and where many plant species fix nitrogen from the air. However, the effect is strong in cooler regions.

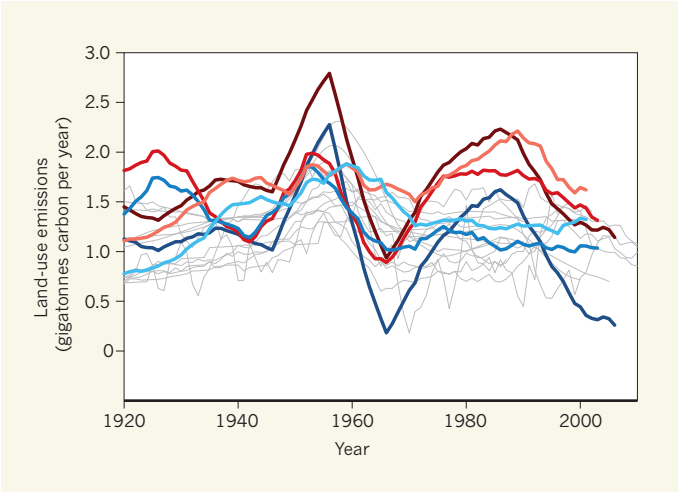


Figure 1 | Land-use emissions and nutrient limitation. Jain *et al.*[1] simulated the emissions that were associated with land-use change using one biosphere model, but three different data sets of land cover by natural vegetation and by managed ecosystems. For each land-cover data set, they performed simulations that included (red lines) or excluded (blue lines) the effects of nitrogen limitation. Each data set's pair of simulations is indicated in equivalent shades (light, medium and dark) of red and blue. The authors find that simulated global land-use emissions are higher when nitrogen limitation is included. Grey lines depict 13 recent estimates of land-use emissions[3], many of which do not account for nitrogen limitation; these illustrate the large range of other uncertainties associated with models of land-use emissions.

Jain and colleagues' emissions estimates that disregard nitrogen limitation are within the range of previous studies, but their estimates that consider nitrogen limitation are at or beyond the high end of those of earlier reports[3] (Fig. 1). If additional studies support the authors' hypothesis, then nutrient cycles will become another source of uncertainty for land-use emissions estimates, the effect of which is about as large as sources that have already been identified — for example, the authors' analysis also shows that estimates of land-use emissions are strongly affected by assumptions made about the extent of land covered by natural vegetation and by managed ecosystems. The choice of biosphere model used and the assumptions made about the carbon density of vegetation are equally important[3].

If historical emissions have been substantially underestimated, for whatever reason, this could be bad news for those who are hoping that

the effects of human activities on climate will be modest. Many climate models have started to integrate processes such as nutrient cycles into their biosphere component, but most of the simulations considered in the upcoming Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) have not. The projections of socio-economic behaviour used in the IPCC report differ widely, but they all indicate that areas containing pristine vegetation will diminish, and that wood harvesting will increase, partly to make biofuels[4]. Future land-use emissions might have been underestimated, and therefore so might the human impact on global climate. Ambitious targets, such as limiting global warming to less than 2 °C above pre-industrial conditions, could be even harder to achieve than was anticipated.

There is also good news, however. Although observational data provide the net exchange of carbon between land and atmosphere, models are needed to work out the contributions made to this exchange by land-use emissions and by the carbon sink that is attributable to environmental changes. These models reproduce the measured net carbon flux, and so if carbon sources have been underestimated then the sink must also have been underestimated. The biosphere might therefore absorb more CO_2 from the atmosphere than is simulated by current models. Further research is needed to identify the processes that contribute to this larger sink, because the efforts required to mitigate climate change directly depend on the size and persistence of the sink in the future.

The study by Jain and co-workers is not the first to consider both nitrogen and land use[5,6], although it is the first to explicitly quantify the effect of nitrogen limitation on land-use emissions. It should be noted, however, that even representations of nitrogen alone in biosphere models are inherently uncertain[2]. In the tropics, the availability of phosphorus could be the more relevant limitation[2]. If nutrient supply is a strong limiting factor to regrowth today, then bookkeeping models that have quantified land-use emissions on the basis of observed changes in carbon stocks[7] will also have captured the effects that are reported by Jain and colleagues. Nevertheless, the new study clearly demonstrates the need to account for 'compound' human disturbances by considering combinations of several factors, such as the rise in atmospheric CO_2, climate change, land use and nutrient cycles.

Even though the authors' estimates of land-use emissions for recent decades (1.7 ± 0.2 gigatonnes of carbon per year for the 1990s, for example) are higher than previous estimates[3] (1.12 ± 0.25 Gt C yr^{-1} for the 1990s), they clearly take second place as a driver of climate change compared with fossil-fuel emissions (6.1 ± 0.3 Gt C in 1990, rising to 9.5 ± 0.5 Gt C in 2011)[8]. But although fossil fuels can be replaced, at some cost, by carbon-neutral alternatives, we will continue to depend on managing Earth's land surface for food and fibre. Any study that helps to identify the relevant processes and to reduce the uncertainties of land-use fluxes is, therefore, a welcome contribution. ■

Julia Pongratz *is at the Max Planck Institute for Meteorology, D-20146 Hamburg, Germany.*
e-mail: julia.pongratz@zmaw.de

1. Jain, A. K., Meiyappan, P., Song, Y. & House, J. I. *Glob. Change Biol.* http://dx.doi.org/10.1111/gcb.12207 (2013).
2. Friedlingstein, P. & Prentice, I. C. *Curr. Opin. Environ. Sustain.* **2,** 251–257 (2010).
3. Houghton, R. A. *et al. Biogeosciences* **9,** 5125–5142 (2012).
4. Hurtt, G. *et al. Clim. Change* **109,** 117–161 (2011).
5. Churkina, G., Trusilova, K., Vetter, M. & Dentener, F. *Carbon Balance Mgmt* **2,** 5 (2007).
6. Zaehle, S., Ciais, P., Friend, A. D. & Prieur, V. *Nature Geosci.* **4,** 601–605 (2011).
7. Houghton, R. A. *et al. Ecol. Monogr.* **53,** 235–262 (1983).
8. Peters, G. P. *et al. Nature Clim. Change* **3,** 4–6 (2013).

GENOMICS

A gut prediction

Characteristic profiles of gut microorganisms in people with type 2 diabetes could aid diagnostics and therapies, but differing signatures between ethnicities and genders highlight the need for further studies. SEE LETTER P.99

WILLEM M. DE VOS & MAX NIEUWDORP

Microbial cells make up the majority of cells in the human body, and most of these reside in the intestinal tract[1,2]. Researchers have long recognized that some intestinal microorganisms are associated with health, but the beneficial impact of most of the gut's microbes on human metabolism has been discovered only relatively recently[2]. It is of great medical and societal importance to pinpoint the associations of these intestinal microbes with health and with diseases such as obesity, metabolic syndrome and type 2 diabetes (T2D)[3]. A study by Karlsson *et al.*[4] on page 99 of this issue is an important contribution to this growing body of evidence*.

Experiments in mice have revealed a causal relationship between certain intestinal microorganisms and obesity[5], and evidence from work in humans suggests[6] that intestinal microbes have a causal role in mitigating insulin resistance — the hallmark of T2D. However, the human intestinal microbiota is immensely complex and includes thousands of species that have a collective genome, termed the metagenome, of close to 5 million genes[4,7,8]. High-throughput sequencing

*This article and the paper under discussion[4] were published online on 29 May 2013.

TABLE 1 | STUDY COMPARISON
A comparison of two metagenome-wide association studies of the gut microbiota of patients with type 2 diabetes.

	Karlsson *et al.*[4]	Qi *et al.*[8]
Population	145 European females.	345 Chinese males and females.
Cohorts	Normal, impaired glucose metabolism and T2D.	Normal and T2D.
Confounding factors	Only postmenopausal females. Some participants on medication.	More males than females. Cohorts not age-matched. Some participants on medication (not reported but probable).
Main findings	Butyrate-producing *Roseburia* species and *Faecalibacterium prauznitzii* lower in T2D cohort. *Lactobacillus gasseri* and *Streptococcus mutans* higher in T2D cohort. Genes involved in oxidative stress response higher in T2D cohort. Genes involved in riboflavin metabolism lower in T2D cohort. Genes involved in assembly of bacterial flagella lower in T2D cohort.	Butyrate-producing *Roseburia intestinalis* and *F. prauznitzii* lower in T2D cohort. Proteobacteria higher in T2D cohort. Genes involved in cofactors and vitamins lower in T2D cohort. Genes involved in assembly of bacterial flagella lower in T2D cohort.

T2D, type 2 diabetes.

of this metagenome, which is derived from stool samples, is now regularly used to analyse the intestinal microbiota and is replacing characterization of individual microbes. When metagenomic analysis is combined with clinical data, it is known as a metagenome-wide association study (MGWAS).

Karlsson and colleagues conducted a detailed MGWAS in 145 European women who had T2D or impaired glucose metabolism (an indicator of pre-T2D), or who were healthy. The authors' results complement those of a similar study, by Qin and colleagues[8], in a group of Chinese men and women that comprised T2D and healthy cohorts (Table 1). The results of both studies are astonishing, showing highly significant correlations of specific intestinal microbes and their genes with T2D. These findings could take approaches for early diagnosis and treatment of T2D far beyond what is possible with existing methods. Moreover, the findings indicate that the predictive power of MGWASs surpasses that of genome-wide association studies, which include only human genes; this is testament to the fact that microbial genes, in addition to our genes, are intricately related to the pathogenesis of T2D, and almost certainly other diseases.

Metagenome analysis is a rapidly emerging field and new sequencing methods and computational-analysis approaches are being developed. Karlsson *et al.* and Qin *et al.* used the same high-throughput sequencing platform, which generates sequences of about 100 nucleotides in size, but their subsequent methods differed: Qin and coworkers mapped their sequences to known metagenome data sets, whereas Karlsson and colleagues constructed their own sequence assembly that they then annotated using newly developed algorithms. Both teams then identified sequences that act as signatures of groups of correlated genes, which Qin *et al.* termed metagenomic linkage groups and Karlsson *et al.* called metagenomic clusters.

When Karlsson *et al.* compared the frequencies of these signatures in T2D patient and control groups, they found that the presence of fewer Clostridiales bacteria that produce the short-chain fatty acid butyrate (*Roseburia* species and *Faecalibacterium prausnitzii*) was highly discriminant of T2D — as had also been seen by Qin and colleagues. Thus, the two studies underscore the known role of butyrate-producing bacteria as regulators of human glucose and lipid metabolism[3], and the function of butyrate (together with other short-chain fatty acids) in maintaining intestinal integrity. However, several associations differed between the two studies (Table 1). For example, Karlsson and colleagues identified an enrichment of *Lactobacillus gasseri* and *Streptococcus mutans*, usually found in the mouth and upper intestinal tract, in their T2D cohort, whereas Qin *et al.* saw an enrichment of Proteobacteria, which may produce inflammatory lipopolysaccharides that lead to endotoxaemia.

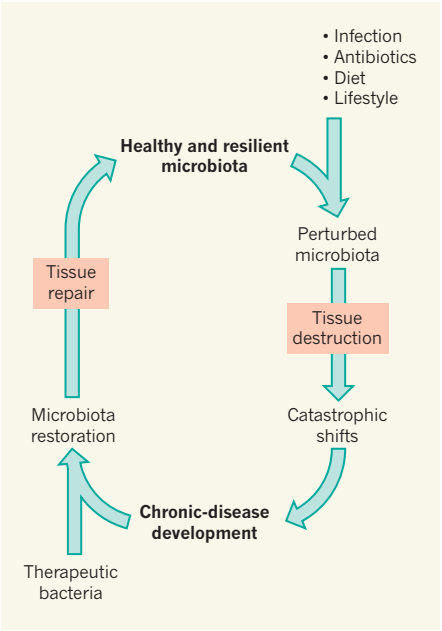


Figure 1 | Microbiota in health and disease. Studies such as that by Karlsson *et al.*[4] contribute to a model of how the composition of gut microorganisms can influence the health of an individual. The model proposes that external factors such as infection or diet alter the healthy, resilient microbial composition to form one that shows early warning signals, such as a reduced number of butyrate-producing bacteria. This altered microbial activity can cause tissue destruction. Further progression can lead to catastrophic composition shifts and chronic disease. If this occurs, it is possible that healthy microbial diversity can be restored, and damaged tissue repaired, only by delivery of specific 'therapeutic' bacteria.

Microbial characteristics such as these might constitute early warning signals[9] that indicate new ecological states in the intestinal microbiota that deviate from the resilient microbiota of healthy people[1,2]. The findings lend support to a model (Fig. 1) in which changes in the composition of gut microorganisms are followed by tissue destruction; in the case of T2D, this affects insulin sensitivity of the liver and muscles. These initial stages could progress to catastrophic shifts in the microbiota that are associated with chronic disease and that may be reversible only through therapeutic intervention to change the intestinal microbiota.

However, further work is needed before our understanding of altered intestinal microbiota can be used in a diagnostic or therapeutic setting. Differences in study design and some confounding factors between the two existing studies (Table 1) could explain some of the differences in outcomes, but long-term, prospective and multi-ethnic cohort studies are needed to determine whether there are in fact substantive differences in predictors of gut microbiota between ethnic groups. Such studies will also need to take diet and gender into account, because both can influence the composition of the gut microbiota[10,11]. Furthermore, for early warning signs (Fig. 1) to be used in a diagnostic capacity, they will need to be shown to have consistent associations with other biomarkers, such as fasting glucose or cholesterol levels, and to be detectable at an early stage in otherwise healthy individuals who are at risk of developing T2D[10].

Despite the need for corroboration of the findings, the potential value of such approaches is underscored by Karlsson and colleagues' demonstration of predictive capacity — the model based on the metagenomic characteristics of their T2D cohort was able to identify women in the pre-T2D cohort who also had high levels of blood-plasma markers associated with T2D. Moreover, such studies might lay the foundation for designing therapeutic bacteria that can be used to 'reset' the intestinal microbiota to the composition that is characteristic of healthy individuals[1,2]. Support for this therapeutic potential comes from microbiota 'transplantations' in patients with T2D and infections of the bacterium *Clostridium difficile* that were able to (temporarily) reshape the composition of the gut microbiota with concomitant beneficial metabolic effects[6,12]. Future strategies might include oral administration of certain intestinal strains as a personalized therapy to postpone or even cure T2D by improving metabolic control in patients. Many of these strains have already been cultured and characterized, so the goal of manipulating our microbiota to keep disease at bay could be closer than we think. ■

Willem M. de Vos *is in the Laboratory of Microbiology, Wageningen University, 6703 HB Wageningen, the Netherlands, and the Departments of Veterinary Biosciences and Bacteriology & Immunology, Helsinki University, Finland.* **Max Nieuwdorp** *is in the Departments of Internal Medicine and Vascular Medicine, Amsterdam Medical Centre, 1105 AZ Amsterdam, the Netherlands.*
e-mails: willem.devos@wur.nl; m.nieuwdorp@amc.uva.nl

1. Lozupone, C. A. *et al. Nature* **489,** 220–230 (2012).
2. de Vos, W. M. & de Vos, E. A. J. *Nutr. Rev.* **70,** S45–S56 (2012).
3. Vrieze, A. *et al. Diabetologia* **53,** 606–613 (2010).
4. Karlsson, F. H. *et al. Nature* **498,** 99–103 (2013).
5. Turnbaugh, P. *et al. Nature* **444,** 1027–1031 (2006).
6. Vrieze, A. *et al. Gastroenterology* **143,** 913–916 (2012).
7. Qin, J. *et al. Nature* **464,** 59–65 (2010).
8. Qin, J. *et al. Nature* **490,** 55–60 (2012).
9. Scheffer, M. *et al. Nature* **461,** 53–59 (2009).
10. Morrow, D. A. *et al. Circulation* **115,** 949–952 (2007).
11. Markle, J. G. *et al. Science* **339,** 1084–1088 (2013).
12. Van Nood, E. *et al. N. Engl. J. Med.* **368,** 407–415 (2013).

REVIEW

doi:10.1038/nature12238

Ice–sheet mass balance and climate change

Edward Hanna[1], Francisco J. Navarro[2], Frank Pattyn[3], Catia M. Domingues[4], Xavier Fettweis[5], Erik R. Ivins[6], Robert J. Nicholls[7], Catherine Ritz[8], Ben Smith[9], Slawek Tulaczyk[10], Pippa L. Whitehouse[11] & H. Jay Zwally[12]

Since the 2007 Intergovernmental Panel on Climate Change Fourth Assessment Report, new observations of ice–sheet mass balance and improved computer simulations of ice–sheet response to continuing climate change have been published. Whereas Greenland is losing ice mass at an increasing pace, current Antarctic ice loss is likely to be less than some recently published estimates. It remains unclear whether East Antarctica has been gaining or losing ice mass over the past 20 years, and uncertainties in ice–mass change for West Antarctica and the Antarctic Peninsula remain large. We discuss the past six years of progress and examine the key problems that remain.

This Review aims to synthesize the main advances in monitoring and modelling of ice-sheet mass balance since the publication of the 2007 Intergovernmental Panel on Climate Change Fourth Assessment Report[1] (IPCC AR4). Mass balance is defined as the net result of mass gains (primarily snow accumulation) and mass losses (primarily meltwater runoff and solid ice dynamical discharge across the grounding line). Surface mass balance (SMB) is the net balance of mass gains and losses at the ice-sheet surface and does not include dynamical mass loss. Efforts to determine ice-sheet mass balance using the three satellite geodetic techniques of altimetry, interferometry and gravimetry (see next section) have recently been sharpened by carefully defining common spatial and temporal domains for inter-comparison[2]. Here we review the latest mass-balance estimates for the Antarctic Ice Sheet (AIS) and the Greenland Ice Sheet (GIS). New glacial isostatic adjustment (GIA) models, tested and evaluated against Global Positioning System (GPS) data, have recently led to significant downwards revision in GIA, and hence downwards revisions of gravimetric and altimetric satellite estimates of Antarctic mass loss[2] (Box 1).

Since the publication of IPCC AR4[1], ice-sheet models are no longer constrained to use overly simplified physics, allowing them to simulate more accurately the important coupling between ice sheets, ice streams and ice shelves. This major advance has been accompanied by improved model representation of the complex interactions of the ice sheet with its bed, the atmosphere and the ocean. For completeness, we also discuss briefly the contributions to sea-level rise (SLR) from other sources, namely glaciers and ice caps, thermal expansion of the oceans and terrestrial water storage changes. Despite recent advances, improved observations and predictions of ice-sheet response to climate change are as urgently needed to feed into mitigation and adaptation models of ensuing SLR as they were at the time of ref. 1.

Recent changes in ice–sheet mass balance

Comparison of mass–balance estimates

One of the most sought after but elusive goals in contemporary Earth science is to relate the mass-balance state of the great ice sheets to observed SLR. A measure of this state provides an unambiguous quantification of the ice-sheet system response to climate change. Recent mass-change estimates have been derived from three categories of techniques.

Volumetric techniques. These determine changes in the volume of the ice sheet via measurements of the height of the ice-sheet surface. They are based on radar altimetry[3,4] or laser altimetry[5].

Space gravimetric techniques. These derive changes in ice-sheet mass via repeated and very accurate measurement of the Earth's gravity field by the Gravity Recovery and Climate Experiment (GRACE) satellite system[6].

BOX 1

Recent developments in GIA models

Glacial isostatic adjustment (GIA) is the response of the solid Earth, including associated changes in planetary gravity and rotation, to past redistributions of ice and ocean mass[92,93]. The clearest observable effect of GIA is regional vertical rebound of the Earth's surface. Models of GIA are necessary for correcting measurements of present-day ice-mass change[94] and for long-term modelling[42]. The assimilation of glacial geological constraints on former ice extent and geodetic constraints on rebound into GIA models is helping to reduce the uncertainty associated with GIA, and hence estimates of ice-mass change[11,12,95]. However, several key challenges remain. First, changes in ice extent and thickness during the past millennium are poorly known, and typically not included in GIA models, despite the fact that they can dominate the present-day rebound signal, especially in regions of low mantle viscosity[96,97]. Second, lateral variations in Earth structure, as detected beneath Antarctica[98], also influence the GIA signal, but are not yet included in most models. Last, the limitations of the data used to tune GIA models mean that probabilistic approaches are now being adopted to seek the most likely range of solutions[99].

[1]Department of Geography, University of Sheffield, Sheffield S10 2TN, UK. [2]Departamento de Matemática Aplicada a las Tecnologías de la Información, Universidad Politécnica de Madrid, 28040 Madrid, Spain. [3]Laboratoire de Glaciologie, Université Libre de Bruxelles, B-1050 Brussels, Belgium. [4]Antarctic Climate and Ecosystems Cooperative Research Centre, University of Tasmania, Aspendale, Victoria 3195, Australia. [5]Department of Geography, University of Liège, 4000 Liège, Belgium. [6]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109-8099, USA. [7]Faculty of Engineering and the Environment, University of Southampton, Southampton SO17 1BJ, UK. [8]Laboratoire de Glaciologie et Géophysique de l'Environnement, UJF – Grenoble 1/CNRS, 38402 Saint-Martin d'Heres, France. [9]Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, Washington 98105, USA. [10]Department of Earth and Planetary Sciences, University of California, Santa Cruz, California 95064, USA. [11]Department of Geography, Durham University, Durham DH1 3LE, UK. [12]NASA Goddard Space Flight Center, Cryospheric Sciences Laboratory, Greenbelt, Maryland 20771, USA.

Mass budget technique. This compares estimates of the net ice accumulation on the ice sheets with estimates of discharge across the grounding line[7] (Box 2).

Each estimate relies on observational data that are unique to its own strategy, and each strategy, therefore, has a unique set of sensitivities to the errors and biases in its data. For example, mass budget[7,8] studies use modelled snowfall fields from atmospheric reanalysis data[9,10] to estimate the mass input into glacier basins, whereas radar and laser altimetry studies use the same fields to estimate the effective density of measured volume changes. Thus mass budget estimates have a first-order sensitivity to errors in the modelled mean accumulation rate, while radar and laser altimetry estimates have only limited sensitivity to errors in fluctuations in the accumulation rate.

Similarly, GRACE and radar and laser altimetry studies require the effects of GIA-related vertical bedrock motion (Box 1) to be removed accurately. Such vertical motion could be misinterpreted as ice-mass change by the GRACE satellites or as ice-thickness change by radar and laser altimeters, and a GIA correction must therefore be applied. This correction is a small percentage (~5%) of the total elevation change typically measured by altimeters; however, the GIA correction applied to GRACE data can be of the same order of magnitude as the signal due to contemporary ice-mass change (because of the density contrast between ice and the solid Earth). As a result, ambiguities in the GIA correction dominate GRACE sources of error in Antarctica (this is not as much of a problem for Greenland where the GIA correction is a much smaller fraction of the total mass change)[6]. Accurate quantification of the GIA signal is therefore crucial; small differences between models can alter the sign of the ice-mass change deduced from GRACE for individual drainage basins[11].

Published estimates of rates of Greenland and Antarctic ice-sheet mass change obtained using the above methods show a large spread of values for the past two decades (Fig. 1). Some of this spread is due to technical differences and some is due to different measurement epochs. However, in the past year, estimates have begun to give a more coherent picture for both Antarctica and Greenland. For Greenland, the trend of increasing mass loss (due to both SMB decrease and ice-to-ocean discharge increase) is clear, while some of the large mass loss estimates for Antarctica have been discarded. We describe some of the improvements in techniques and analysis below.

Reduced uncertainties

Recent assessments of mass-balance history[12,13], coupled with more robust GPS observations of the motion of exposed bedrock[14], strongly suggest that Antarctic GIA-related bedrock motion peaks at about 5–6 mm yr^{-1}. The resulting GIA models for Antarctica[13,15] deliver less than half the mass corrections implied by previous models. At the same time, processing of GRACE data has become more consistent between groups as the time series lengthens. Estimates using the latest models show moderate, if increasing, decadal mass losses for Antarctica[13,16,17].

In the IMBIE (Ice-sheet Mass Balance Inter-comparison Exercise) project, researchers recently compiled average sets of mass-balance estimates for common time periods for both the Antarctic and Greenland ice sheets, using the latest data, with multiple groups deriving estimates with each technique[2]. An important technical change helped reduce the difference among techniques: unlike previously published mass budget estimates that extrapolated mass changes from surveyed to unsurveyed basins, the IMBIE mass budget estimates use radar altimetry data to demonstrate that unsurveyed areas have near-zero rates of mass change, giving, on average, less mass loss. Other extrapolation techniques can give a more positive Antarctic balance for the same data[18]. Similarly, including the most recent GIA estimates for Antarctica brought GRACE estimates closer to the radar and laser altimetry estimates. The IMBIE estimates are simple averages of all measurements, and the discordance that remains among methods (between radar and laser altimetry, for example) is not fully understood.

Figure 1 shows that the disparity of recent mass-balance results among different techniques—primarily from IMBIE—is considerably reduced from that seen before. There tend to be systematic differences between the results from different techniques, with the mass budget method giving the most negative estimate for both ice sheets, laser altimetry the most positive, and GRACE in between. IMBIE radar altimetry estimates cover only the sub-peninsular part of Antarctica, and give rates of mass change consistent with those from GRACE. The techniques agree in sign, and roughly in magnitude, for Greenland, and there is considerable basin-scale spatial fidelity revealed in the inter-comparisons. Greenland had small contributions to SLR in the 1990s (-51 ± 65 Gt yr^{-1}) but was recently (2005–10) losing mass at -263 ± 30 Gt yr^{-1} (ref. 2). (We note that 362.5 Gt yr^{-1} = 1 mm yr^{-1} sea-level equivalent.) The situation for Antarctica is less clear, with one estimate showing a significant positive mass balance[19]. An unweighted average of the estimates indicates that Antarctica, which was in a state of weakly negative balance in the 1990s, is now losing mass at a rate between -45 and -120 Gt yr^{-1}, with large dynamic losses in West Antarctica partially offset by SMB gains in East Antarctica.

For Greenland, an independent group of researchers compared laser altimetry, mass budget and GRACE estimates over the 2003–09 ICESat

BOX 2
Grounding lines and buttressing

Marine ice sheets, such as the West AIS, rest on bedrock that lies below sea level. These grounded ice sheets are fringed by floating ice shelves. The grounding line is the contact of the ice sheet with the ocean where the ice mass starts to float by buoyancy. Ice from the grounded ice sheet is discharged across the grounding line into ice shelves, from where icebergs break off, through a process called calving (Fig. 3).

The migration of the grounding line is a result of the local balance between the masses of ice and displaced ocean water. The grounding line advances if previously floating ice becomes thick enough to ground, or retreats if previously grounded ice becomes thin enough to float. Theory has demonstrated that in order to simulate grounding-line migration, it is necessary to include (horizontal) stress gradients across the grounding zone[22] and in order to resolve this numerically, a high spatial resolution is needed, either by using a moving grid (following the grounding line directly) or by subsampling the grid around the grounding line to hundreds of metres (ref. 39). This high resolution is necessary to resolve horizontal stress gradients across a narrow boundary layer.

Ice discharge generally increases with increasing ice thickness at the grounding line. For a bed sloping down towards the interior this may lead to unstable grounding-line retreat, as increased flux (for example, due to reduced buttressing) leads to thinning and eventually flotation, which moves the grounding line into deeper water where the ice is thicker. Thicker ice results in increased ice flux, which further thins (and eventually floats) the ice, which results in further retreat into deeper water (and thicker ice), and so on (Fig. 3). This unstable retreat is referred to as the marine ice-sheet instability[22]. However, the grounding line is partially stabilized by the presence of ice shelves, which are either confined laterally through embayments or otherwise stabilized by locally grounded features which they enclose (for example, pinning points). Both geometries transmit a back-force, or 'buttressing', towards the grounded ice sheet, which may help to stabilize the grounding line against unstable retreat down inland-sloping bedrock[100].

Thinning of ice shelves reduces drag at the margins and over pinning points, leading to increased ice flow across the grounding line, causing grounding-line retreat until a new stable point (for example, upward sloping bedrock) is reached. The mechanisms described above rely heavily on a precise knowledge of the geometry of the ice–ocean contact, which explains why neighbouring outlet glaciers, in contact with the ocean, and subject to the same atmospheric and oceanic forcing, may exhibit contrasting behaviours[30].

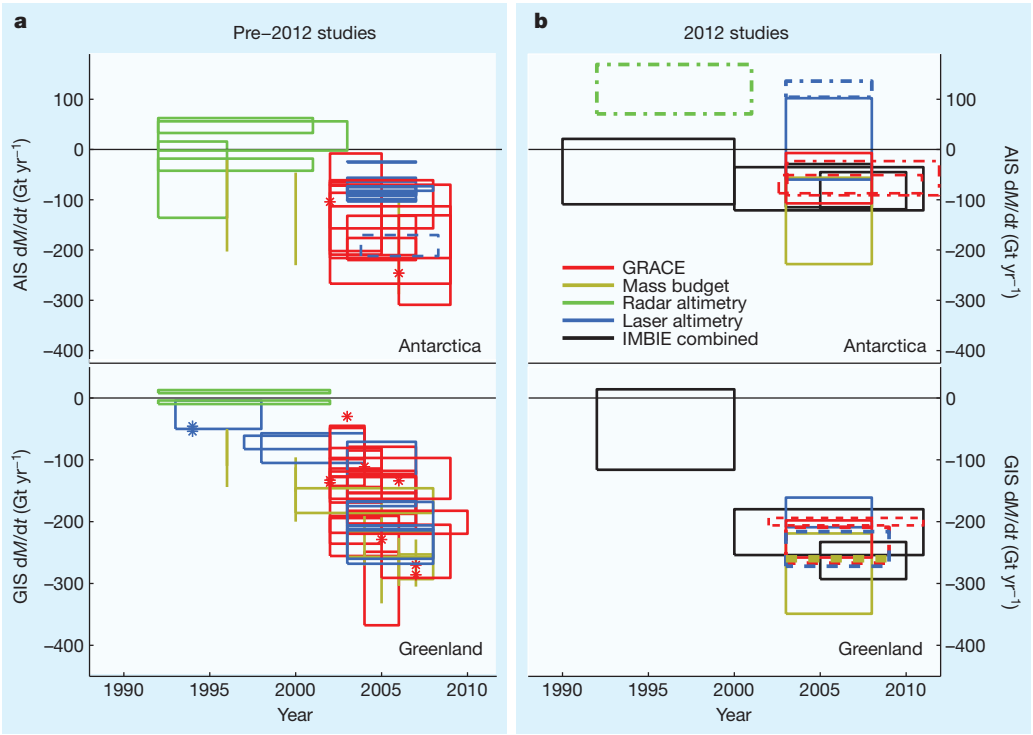


Figure 1 | Summary of estimates of rates of ice mass change for Antarctica and Greenland. In the studies published before 2012 (ref. 2, **a**) and in 2012 (**b**), each estimate of a temporally averaged rate of mass change is represented by a box whose width indicates the time period studied, and whose height indicates the error estimate. Single-epoch (snapshot) estimates of mass balance are represented by vertical error bars when error estimates are available, and are otherwise represented by asterisks. Line colour indicates mass assessment technique (see key); line type indicates data source. 2012 studies in **b** comprise IMBIE combined estimates[2] (solid lines), and estimates by Sasgen and others[16,20] and King and others[11] (dashed lines), Zwally and others[19] (dot-dashed lines), Harig and Simons[89] and Ewert and others[90] (dotted lines).

(Ice, Cloud, and land Elevation Satellite) period: the mass budget estimate gave the maximum loss rates at -260 ± 53 Gt yr^{-1} and GRACE the minimum, at -238 ± 29 Gt yr^{-1} (ref. 20). On a basin-by-basin basis, agreement between the mass budget method and other techniques provides validation for the practice of partitioning mass-balance change between discharge and SMB components, demonstrating that in the northern part of Greenland, the dominant cause of mass change was atmospheric in origin, while in the southern part it was ice dynamics.

The new, reconciled IMBIE GRACE estimates of whole Antarctic mass balance are now largely in agreement with one another, with spreads of 30–50 Gt yr^{-1} between the largest and smallest 2003–08 rates. Previously published GRACE values show spreads around twice as large for similar time periods. In the Antarctic Peninsula and West Antarctica, the IMBIE estimates from laser altimetry and GRACE are in good agreement, in contrast to East Antarctica[2]. For East Antarctica, a mass gain of $+101$ Gt yr^{-1} for 2003–08 has been proposed recently on the basis of laser altimetry[19], which is larger than the IMBIE GRACE estimate of $+35$ Gt yr^{-1} and near the upper end of the laser altimetry estimates[2].

Recent advances in ice-sheet modelling

Key improvements and future challenges

Significant improvements in ice-sheet modelling have been made since the publication of IPCC AR4[1], motivated by the need to understand continuing changes and by the challenge to make more realistic projections for the next few centuries. The primary improvements concern mechanical approximations made to the ice flow equations. The very first generation of ice-sheet models was based on the shallow ice approximation[21]. Such models assume that all resistance to flow is provided by shear-stress gradients in the vertical, which is valid for creeping ice-sheet flow, but not when other ice-dynamical features such as ice streams and ice sheet/ice shelf coupling come into play in ice-sheet evolution. More recent ice-sheet models now include horizontal stress gradients, and can be classified into four categories of increasing complexity and computational cost. (1) Ice shelf/stream models are based on the shallow-shelf approximation[22]. They include horizontal stress gradients, but neglect the vertical shear stresses (which is valid for rapid ice flow at low basal traction). (2) Hybrid models use some combination of solutions from the shallow-ice approximation (to account for the vertical shearing component of flow within grounded ice) and the shallow-shelf approximation (to account for the horizontal stress coupling taking place in ice shelves or regions of rapid sliding)[21,23,24]. (3) More elaborate higher-order models treat the vertical dimension more rigorously, with the only approximation being the hydrostatic assumption (pressure at any point in the ice is due only to the weight of the ice above it and not due to ice flow)[25,26]. (4) Finally, a few models solve the equations of motion without neglecting any terms. These are called 'Full Stokes models', and have recently demonstrated their ability to perform century-timescale simulations applied to a whole ice sheet[27,28].

Spatial resolution of models is the second aspect that has been improved. Hardly any model is now run with a spatial grid size greater than 20 km, but this resolution is still not high enough to resolve ice streams, which are often only a few kilometres wide. Moreover, grounding-line migration and calving require subkilometre resolution. Unstructured grids (for finite element models[27,28]) or adaptive mesh refinement[29] are two strategies that have proven efficient at treating this difficulty with acceptable computational cost.

A third improvement has been enabled through satellite and ground-based observations, such as the quantification of surface velocities and velocity change from satellite interferometry[30], surface elevation change through satellite and airborne campaigns (IceBridge), and high-resolution bedrock and ice thickness measurements[31]. Ice-sheet model behaviour is highly dependent on initial and boundary conditions and faces the difficulty that drag at the ice–bed interface is poorly known. Inverse methods

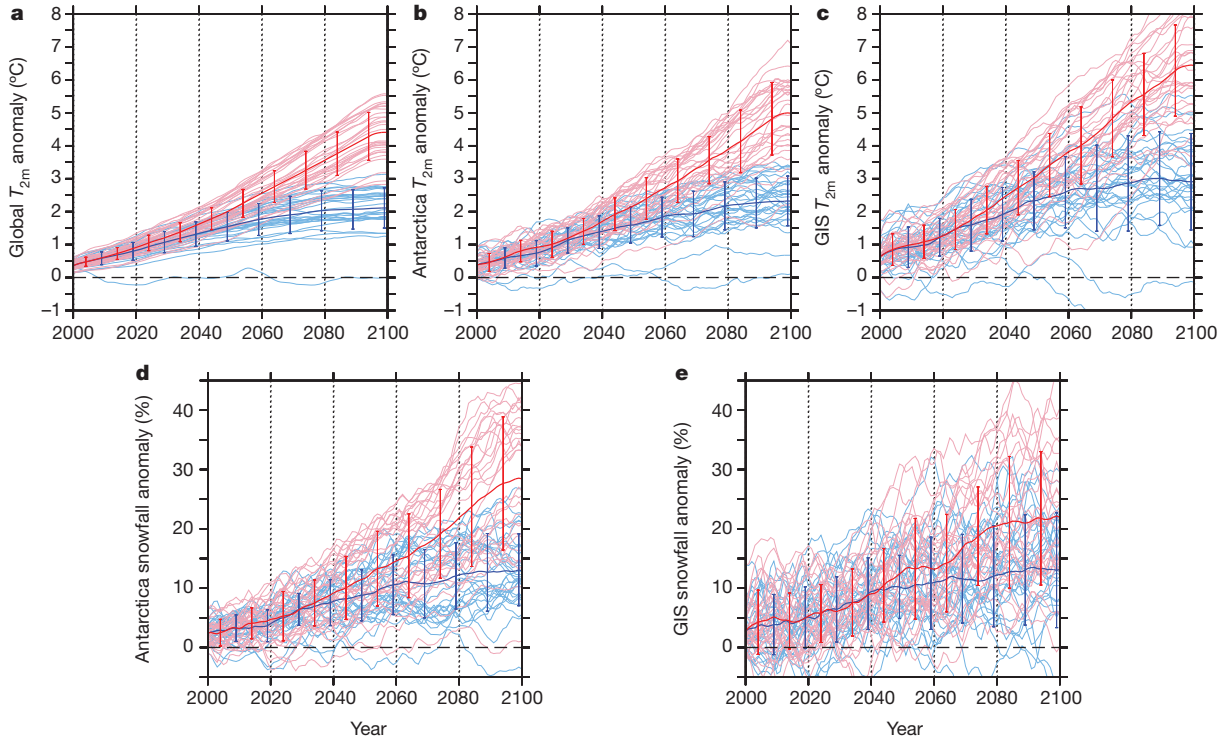


Figure 2 | Comparison of projected global, Antarctic and Greenland surface air temperature and snowfall anomalies to 2100. **a**, Anomaly of global mean 2 m air temperature (T_{2m}) simulated by 30 GCMs from the CMIP5 data base. Values are with respect to 1970–99 for the RCP 4.5 (blue) and RCP 8.5 (red) scenarios. We refer to ref. 91 for more details about the Representative Concentration Pathways (RCP) scenarios. The evolving ensemble means are plotted as thick lines, with vertical bars representing ±1s.d. for each decade. A 10-year running mean was used to smooth the curves. **b**, Same as **a** but for Antarctica. The land/sea mask from each GCM is used to delimit Antarctica. **c**, Same as **a** but for T_{2m} over GIS. The T_{2m} anomaly is taken over the area covering Greenland (60–85° N and 20–70° W) and where surface elevation is higher than 1,000 m above sea level. **d**, Same as **b** but for precipitation. Anomalies are given in per cent with respect to the mean precipitation for 1970–99. **e**, Same as **c** but for precipitation.

have now been successfully implemented in ice-sheet models to infer the basal drag map that provides a good agreement between observed and simulated surface velocities. This procedure is becoming standard in the spin-up that is required for establishing an optimum initial state[27–29,32]. All the above refinements enable models to reproduce present-day observed ice-sheet flow speeds, which is a major improvement since AR4[1] was published.

Grounding lines, sliding and calving

Warming-induced ice-shelf loss has caused major glaciers and ice streams of Antarctica to speed up[33,34]. The mechanisms behind this speed-up are complex. Oceanic and/or atmospheric warming leads to ice-shelf thinning or disintegration[35,36], which in turn may lead to loss of buttressing[37], grounding-line retreat and hence glacier speed-up[33] (Box 2). Observations from the Antarctic Peninsula and the Amundsen Sea Embayment in West Antarctica (for example, Pine Island and Thwaites glaciers, which are currently the main contributors of the AIS to SLR[38]) support these mechanisms.

Major theoretical advances[22] in understanding the motion and stability of the grounding line show that in the absence of buttressing (see Box 2), grounding lines retreat unstably on an upward-sloping bed (in the direction of ice flow). Analytical solutions are now available to test and verify marine ice-sheet models, so that the numerical error associated with predicting grounding-line motion can be reduced significantly to the level of parameter uncertainties[39]: models that attempt to account for grounding-line dynamics should incorporate horizontal stress transmission across the grounding line, so that the grounded ice sheet realistically feels the influence of floating ice (Box 2). Furthermore, the grounding line needs to be resolved at a sufficiently high spatial resolution[39]. Such developments have been made recently and applied to Pine Island glacier, where a small increase in sub-ice-shelf melting has been shown to result in either unstable grounding-line retreat[29], or grounding-line stabilization approximately 25 km inland within 100 years (ref. 37).

GIA also influences ice-sheet behaviour[40,41]. Effects such as Earth's deformation in response to ocean loading, and perturbations to the shape of the sea surface in response to the redistribution of both internal and surface masses, including changes to the mass of the ice sheet itself, play a key role in governing the behaviour of a marine-grounded ice sheet, such as West Antarctica[42]. GIA alters the water depth via spatially varying perturbations to both the ocean floor and the sea surface and this has a first-order effect on grounding-line positions[22]. Ignoring such processes can fundamentally alter model predictions relating to the stability of a marine-grounded ice sheet[41].

Ice flow across the grounding line is equally controlled by inland basal hydrological conditions and processes that govern basal sliding and sediment deformation. A wide range of observations over the GIS suggests that surface melt water reaches the bed by fracture and drainage through moulins, and this is likely to affect basal lubrication[43]. Recent work has shown that it is not simply mean surface melt but an increase in water input variability that drives faster ice flow[44]. This has been confirmed by observations[45]. However, more recent work supports the original contention that increased melt water leads to increases in basal sliding, but that the effect is much smaller than originally thought because of buffering by subglacial drainage system evolution[46]. Given the available evidence, the representation of basal sliding in large-scale ice-sheet models still depends largely on empirical parameterizations based on observations of seasonal variations in ice flow.

Recent developments in the understanding of calving follow either fundamental process approaches[47,48], leading to global calving laws relating thickness at the grounding line/calving front to calving rate, or are based on stochastic modelling and fracture theory[49]. Two-dimensional generalizations of similar calving laws have been proposed in large-scale models[50]. More specific approaches take into consideration environmental factors,

relating surface meltwater runoff and sub-shelf melting to the widening of crevasses and subsequent calving[51]. However, model applications based on this approach remain restricted to one-dimensional flowline models[52], owing to the lack of data to resolve the geometry of outlet glacier embayments at sufficiently high spatial resolution. Although improvements have been made over recent years, this lack of data hampers a complete process-based evaluation of calving. In the near future, it is likely that models will continue to rely on empirically based parameterizations of calving.

Future ice-sheet changes

For significantly warmer climates, both the GIS and AIS are projected to lose mass[53]. General circulation models (GCMs) generally project a small increase of snowfall over both ice sheets (Fig. 2d, e). However, the mass loss from increasing surface melt will be dominant over the GIS. For Antarctica, although the SMB is projected to increase, there remain major uncertainties concerning the response of the marine ice sheets and ice shelves to ocean forcing.

Surface melt already occurs over a large part of the GIS during summer and reached a new record in 2012[54]. Therefore, rising temperatures will mainly affect mass loss through increased surface melt in summer, and several positive feedbacks may accelerate this surface mass loss:
(1) Polar amplification of global warming resulting from, among other processes, the decrease of sea-ice extent over the Arctic Ocean and its associated positive albedo feedback. This process, already observed in recent years[55] and simulated by the Coupled Model Inter-comparison Project Phase 5 (CMIP5) GCMs (see Fig. 2c compared with Fig. 2a), doubles the estimated uncertainties in projected near-surface temperature anomalies for Greenland compared with those at the global scale[56].
(2) Positive snow albedo feedback over the ice sheet itself associated with the expansion of the bare ice zone. This effect explains why the meltwater runoff increases quadratically with rising summer temperatures: the albedo of bare ice (0.3–0.5) is much less than that of melting snow (~0.7), and surface melt water becomes more likely to run off rather than percolating into deeper parts of the snowpack[57].
(3) Positive elevation feedbacks associated with the thinning of the ice sheet resulting from the increasing surface melt and ice discharge. Significant thinning (up to 100 m) of the ice sheet is projected along the ice-sheet margin[58], which should cause an additional melt increase over this area (as ice moves to lower elevations, where it is warmer).

Dynamical changes of the GIS due to enhanced lubrication, calving and ocean warming still remain difficult to predict. Higher-order ice flow modelling of observed retreat of GIS glaciers over the past decade and subsequent upscaling (extrapolation of these model results to the whole GIS) leads to a minimum additional SLR of 6 ± 2 mm by 2100, with an upper bound of 45 mm when recurring forcing is applied[59], while similar upscaling of realistic atmospheric and oceanic forcing of four GIS glaciers with a calving model leads to a maximum dynamic contribution of 40–85 mm by 2100[60]. This is still lower than previous estimates, but higher than when this retreat chronology is implemented in a three-dimensional higher-order model, leading to a dynamic contribution of 7–15 mm (ref. 61). The reason for such low numbers is that owing to the retreat of the ice-sheet margin, calving seems to decrease in relative importance[53,61]. According to a model inter-comparison[62], increased ice shelf melt rates of 2 m yr^{-1} lead to 27 mm SLR by 2100 (and 135 mm from a high melt rate of 20 m yr^{-1}). In response to SMB changes, ice-sheet model results are quite consistent and most studies conclude that the largest uncertainty comes from the spread among global climate models, which is amplified by some of the above-mentioned feedbacks over Greenland[56,58].

For Antarctica, the amplification of the global climate modelling uncertainties is smaller and the contribution of Antarctica to SLR is predicted to increase logarithmically with rising global temperatures (as positive feedbacks become increasingly apparent later) but with little change, and even perhaps a negative contribution, in the next 100–200 years (ref. 53). First, polar amplification resulting from reduced sea-ice coverage seems to be smaller than for the Arctic (see Fig. 2b). However, a changing Antarctic Circumpolar Current could potentially allow warmer water to penetrate into the coastal shelf regions of Antarctica—as is observed[63]. Second, little surface melt currently occurs and rising temperatures are not expected to enhance surface melt significantly in the next 100 years (ref. 53). Third, an increase in snowfall is expected to be more significant owing to atmospheric temperature rise, hence leading to an increase in SMB[64]. Here, the elevation feedback resulting from SMB changes is negative because the ice sheet is initially projected to thicken[53], which is expected to affect its dynamics, especially on longer than centennial timescales.

The response of ice-sheet dynamics is twofold, due to increased accumulation and to higher ocean temperatures (in particular below the ice shelves). Two models[53,65] produce ice-sheet thickening over East Antarctica and increased ice flux at the grounding line due to higher snowfall. However, both studies[53,65] fail to account for processes at the ice-sheet/ice shelf/ocean interface, such as grounding-line retreat or loss of buttressing[39]. So far, a continental-scale Antarctic ice-sheet model assessment taking into account those fundamental processes is lacking, although one assessment—based on a wide variety of model complexities—does report large inter-model variability in response to ocean forcing[62]. Process-based modelling of parts of the West AIS, such as Pine Island glacier, results in a SLR contribution of 27 mm by 2100 for a modest grounding-line retreat of 25 km (ref. 37), whereas significant (100 km) grounding-line retreat is modelled elsewhere[29]. An alternative method based on probabilistic extrapolation of sustained glacier retreat from such numerical model output[37] leads to a SLR contribution of 130 mm by 2100[66].

Other contributions to SLR

The global average rate of SLR over the past few decades is about 2–3 mm yr^{-1} (ref. 67). Estimates of the global contribution from glaciers and ice caps (GICs) to SLR in the IPCC AR4[1], 0.50 ± 0.18 mm yr^{-1} (1961–2003) and 0.77 ± 0.22 mm yr^{-1} (1993–2003), were based on extrapolation of sparse mass-balance measurements made by the glaciological method[1] (Box 3). These values were later considered underestimates[68], owing to the poor representation in the glacier inventories of the GICs peripheral to Greenland and Antarctica (peripheral GICs, PGICs): thus the 1961–2003 value was raised, based on a combined modelling and observations approach[68], to 0.79 ± 0.34 mm yr^{-1} (no value was provided for 1993–2003). A later extrapolation-based global estimate[69], with the novelty of allowing explicitly for glacier shrinkage, resulted in a lower estimate of 0.63 mm yr^{-1} for 1961–2006 (no uncertainty was given). The extrapolation-based global estimates have been improved by the addition of geodetic mass balances (Box 3) to the inventories of mass balance calculated using the glaciological method, which has resulted in consistently larger contributions to SLR, especially for the

BOX 3

Glaciological versus geodetic method

GIC mass-balance estimates by the glaciological method are based on extrapolation over the whole glacier surface of measurements of accumulation and ablation made *in situ* at single points. These measurements include readings of surface elevation changes at stakes, sampling of density and accumulation in pits and shallow cores, depth probing of the snow and firn, and shallow coring.

Estimates by the geodetic method are based on repeated mapping of glacier surface elevations to estimate the volume changes, from which the mass changes are calculated using information about the density of the material and its time variations. The elevation changes can be measured using different techniques, either from the glacier surface or, more commonly, from airborne or satellite-borne sensors.

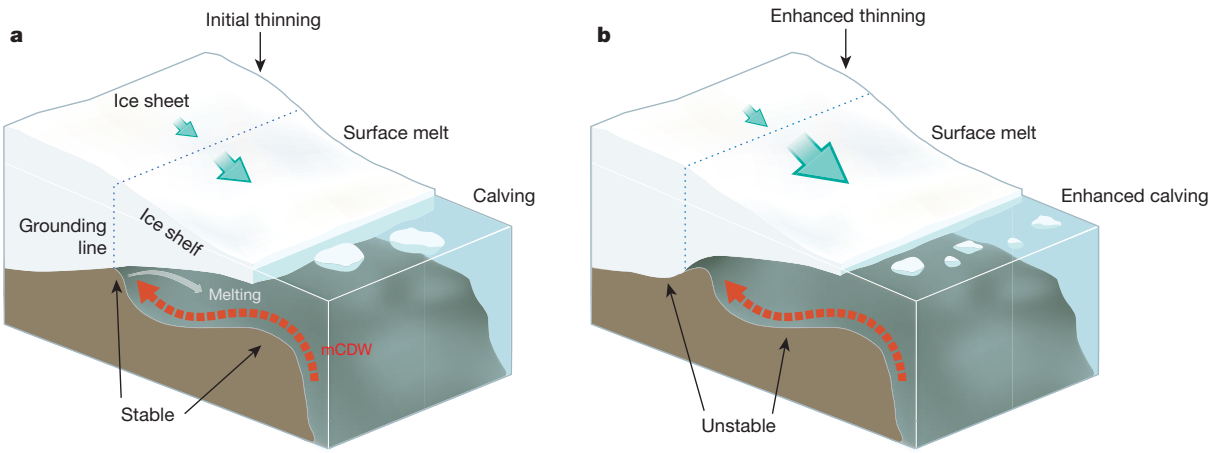


Figure 3 | Illustration of a marine ice sheet and its interaction with the ocean. a, Warm modified Circumpolar Deep Water (mCDW) leads to melting at the grounding line, leading to ice-shelf thinning, grounding-line retreat, and initial thinning. **b**, Marine ice-sheet instability occurs when, in the absence of buttressing, the grounding line retreats on an upward-sloping (in the direction of the flow) bedrock (unstable): ice flux increases with thickness at the grounding line, leading to an increased outflux to the ocean and enhanced thinning that may be compensated by further grounding-line retreat, until a new downward-sloping bed (pinning point) is reached (stable). Thinning of ice sheet and shelf can also be caused by surface melt and increased calving.

most recent periods (for example, 0.99 ± 0.04 and 1.46 ± 0.34 mm yr^{-1} for 1993–2008 and 2000–05, respectively[67,70], compared with 0.97 and 0.95 mm yr^{-1} for 1993–2006 and 2002–06 respectively[69]).

Satellite gravimetry, a method traditionally restricted to the large ice sheets, has recently been used to estimate the global contribution of GICs to SLR[71]. GRACE data alone do not have the resolution to separate the Greenland and Antarctic ice sheets from their PGICs, but using an upscaling approach similar to that of ref. 68 has allowed one group to estimate a global contribution from GICs to SLR of 0.63 ± 0.23 mm yr^{-1} during 2003–10[71], which is 30% and 47% lower than the two previous estimates that most closely match this period (2002–06[69] and 2005–10[72], respectively). GRACE results for GICs, however, are sensitive to the models used for calculating GIA, post-Little Ice Age isostatic rebound, and surface- and ground-water mass transfer corrections.

The large uncertainties associated with the conventional extrapolation-based methods mostly arise from the uneven representation of the glacier-covered regions in the mass-balance measurements and the incomplete knowledge of the PGICs, both in terms of poorly known mass balances and inaccurate estimates of their area. The latter has greatly improved with the recent release of the Randolph Glacier Inventory[73]. A consensus estimate combining GRACE, laser altimetry and the extrapolation-based method, using a common inventory of glaciers and a common spatial and temporal reference[74], has very recently enabled reconciliation of the disparate global estimates of wastage from GICs so far available from the different techniques. The consensus value is 0.71 ± 0.08 mm yr^{-1} during 2003–09, which is far lower than the extrapolation-based approach[72] and somewhat higher than the GRACE-based estimate[71].

Ocean thermal expansion (OTE) is a major component of the SLR observed during the late twentieth century[67], and is projected to continue through the twenty-first century and beyond[75]. The IPCC AR4 found that OTE contributed ~25% of the observed SLR for 1961–2003 and ~50% for 1993–2003[1]. Time-varying biases in the ocean temperature data, however, were recently detected[76] and reduced. It is now understood that the percentages of SLR explained by OTE during the above periods are almost identical[77], and so are higher for 1961–2003 and lower for 1993–2003 than estimated in the IPCC AR4[1]. A recent sea level budget[67] indicates that OTE contributed ~40% of the observed SLR since 1970 and ~30% since 1993. Warming in the upper 700 m of the ocean explains about 70–80% of the OTE rates. Multi-decadal rates[77] for OTE in the upper 700 m are 0.71 ± 0.10 mm yr^{-1} for 1970–2011 and 0.85 ± 0.20 mm yr^{-1} for 1993–2011, based on linear regression and time-variable uncertainties. Multi-decadal rates for the deep/abyssal ocean are very uncertain[67], as these are the most poorly sampled regions of the ocean. Since 2005, about 3,000 autonomous Argo profiling floats have been monitoring the upper 2,000 m of the ocean. The Argo-based OTE rate[78] for 2005–11 is 0.60 ± 0.20 mm yr^{-1}, in close agreement with the change inferred from satellite altimetry and GRACE[79]. Although consistent with the rates estimated for the multi-decadal periods, the OTE rate for 2005–11 is unlikely to represent long-term changes. Over such a short period, long-term changes can be easily obscured by more energetic ocean variability, such as fluctuations in the phase of the El Niño/Southern Oscillation[80].

Recent estimates for total terrestrial water storage changes during 1993–2008, which include dam retention, groundwater depletion and natural terrestrial storage changes, give values ranging from -0.08 ± 0.19 mm yr^{-1} (ref. 67) to 0.10 ± 0.20 mm yr^{-1} (ref. 81). A much larger (positive) contribution dominated by groundwater depletion has recently been suggested[82], although this result is still controversial[83].

Table 1 summarizes the recent and current contributions to SLR calculated with the methods discussed in this Review and compares their sum with the observed SLR from tide gauges and satellite altimetry[67]. OTE appears as the main current contributor to SLR, closely followed by the large ice sheets, whose contribution is increasing, and the GICs. The contribution from land-ice masses (ice sheets and GICs) could be

Table 1 | Estimated recent and current contributions to SLR

Source of contributions	SLR (mm yr^{-1})	
	1992/93 to 2008/11*	2000/03 to 2009/11*
GIS + AIS[2]	0.59 ± 0.20	0.82 ± 0.16
GICs[72,74]	1.40 ± 0.16	0.71 ± 0.08
Ocean thermal expansion[77,87,88]	1.10 ± 0.43	1.11 ± 0.80
Terrestrial water storage (1993–2008)[67,81]	0.02 ± 0.26	
Sum of contributions	3.11 ± 0.56	2.66 ± 0.86
Observed (1993–2008)[67]	3.22 ± 0.41	

For 'Terrestrial water storage' and 'Observed', only the values for the longer time span are given; the terrestrial water storage number is used for the sum of contributions for both periods; 'Observed' means observed SLR from tide gauges and satellite altimetry. For GICs we have taken an update of the values given in ref. 72 for 1993–2011, while for 2003–09 we have used the value given in ref. 74. The value given for 'Ocean thermal expansion' combines a long-term abyssal value[87] with updates, for the periods shown in the table, from an average of refs 77 and 88 for the uppermost 700 m, and from ref. 88 for 0–2,000 m. The value given for terrestrial water storage is an average of those in the references shown. The uncertainties given are the published errors from the individual studies (usually standard deviations). When data from several sources are combined, the quoted error for the sum of contributions is the square root of the sum of the individual variances.

* The two periods given here need to accommodate data from a variety of sources, and so flexible start and finish dates are given. For example, '1992/93 to 2008/11' means that the data in the column below start in 1992 or 1993, and end somewhere between 2008 and 2011.

slightly overestimated, because only some of the methods in the consensus estimate for ice sheets[2] explicitly exclude the PGICs (and thus the contribution from PGICs may have been double-counted). Also, the apparent decrease in the contribution from the GICs between the two periods (Table 1) is mostly a result of the different methods used, rather than a result of a lower SMB observed during 2005–10[72] (to illustrate this, we note that the GICs SLR contribution given in ref. 72 for 2000–10 is 1.38 ± 0.21 mm yr^{-1}). Note that, for the most recent period, there is a gap between the sum of contributions and the SLR observed from tide gauges and satellite altimetry.

Conclusions and outlook

During the past 20 years, the AIS as a whole (East, West and Antarctic Peninsula) has been losing mass, and this is certainly true of the GIS[2]. There are still disagreements between the numbers that come from the mass-balance retrieval techniques, particularly for East Antarctica, demonstrating a need to understand the errors of each method better. For radar altimetry, further assessment is needed of surface-density corrections and of short-term corrections to ENVISat radar altimetry data[84], as more moderate estimates of rates of mass change are possible using such corrections. For the mass budget method, NASA's IceBridge project will provide airborne-radar-based improvements to SMB estimates, and radar-sounding measurements of ice thickness at grounding lines will provide improved discharge estimates. Gravimetry and laser altimetry will have, respectively, GRACE and ICESat-2 follow-on missions (scheduled launches in 2017 and 2016, respectively) that will ideally provide a decadal record of whole ice-sheet mass balance. However, it is unlikely that these refinements will change the consensus picture emerging: whereas Antarctica as a whole is losing mass slowly (assessed to be contributing 0.2 mm yr^{-1} sea-level equivalent by IMBIE[2]), Greenland, the Antarctic Peninsula and parts of West Antarctica are together losing mass at a moderate ($\sim$1 mm yr^{-1} sea-level equivalent) rate today ($\sim$70% of this mass loss is from Greenland) and rates for each are becoming increasingly negative. For the past decade, the collective sea-level contribution from the ice sheets is similar to those from each of GICs and oceanic thermal expansion.

Although the West AIS is most probably going to continue to contribute to SLR (although the amount is poorly constrained), the sign of the contribution of the East AIS over the next century is uncertain. From the standpoint of projecting global sea level through this century and beyond, it is of fundamental importance to focus on improving ice-sheet models, including representation of key processes and nonlinear transitions. The concern of policymakers rightly focuses on the possibility of extreme outcomes, with their large impact potential and adaptation need[85]. This is particularly true for the cryosphere, which responds nonlinearly to rising temperatures because of several potential positive feedbacks that may accelerate deglaciation. Improved knowledge of key ice-sheet thresholds would support climate policy decisions. Continued observations of ice-sheet processes and their implementation in ice-sheet models are crucial to ensure more accurate sea-level projections.

We have identified several important challenges that remain. First, there is a need for upscaling parameterizations to allow low-resolution models, which run fast but with coarse meshes, to represent crucial processes better. So far, parameterizations for grounding-line migration have been proposed[22,23] and tested against more complete models[39]. Although advances have been made on the theoretical level, process-based calving implemented in numerical flow models still has to rely on parameterizations that are not fully verified against physical models. Second, although progress has been achieved in the spin-up of ice-sheet models so that initial states are closer to observations through the use of inversion techniques, the nonlinearity of basal drag and its dependency on basal hydrology remains a concern. Time-dependent evolution of basal drag is not yet fully implemented in operational models, partly because subglacial hydrology models have not yet been fully implemented and partly because the data required to calibrate spatially dependent basal friction laws are lacking. The recent release of velocity maps for various time periods[86] gives hope that this problem will soon be tackled. Third, a further vital step will be to couple improved ice-sheet models with atmosphere/ocean models and GIA models to account for all the feedbacks between the various physical systems at sufficiently high resolution. This will need to be supported by targeted observations with appropriate spatial and temporal coverage.

Received 3 January; accepted 26 April 2013.

1. Solomon, S., *et al. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2007).
2. Shepherd, A. *et al.* A reconciled estimate of ice sheet mass balance. *Science* **338,** 1183–1189 (2012).
   Gives an overall view of remote sensing of ice-sheet mass balance and arrives at a nearly reconciled estimate of the contribution of the ice sheets to sea-level rise.
3. Davis, C. H. & Li, Y. H. McConnell, J. R., Frey, M. M. & Hanna, E. Snowfall-driven growth in East Antarctic ice sheet mitigates recent sea-level rise. *Science* **308,** 1898–1901 (2005).
4. Zwally, H. J. *et al.* Mass changes of the Greenland and Antarctic ice sheets and shelves and contributions to sea-level rise: 1992–2002. *J. Glaciol.* **51,** 509–527 (2005).
5. Zwally, H. J. *et al.* Greenland ice sheet mass balance: distribution of increased mass loss with climate warming. *J. Glaciol.* **57,** 88–102 (2011).
6. Velicogna, I. Increasing rates of ice mass loss from the Greenland and Antarctic ice sheets revealed by GRACE. *Geophys. Res. Lett.* **36,** L19503 (2009).
7. Rignot, E. & Kanagaratnam, P. Changes in the velocity structure of the Greenland ice sheet. *Science* **311,** 986–990 (2006).
8. Rignot, E., Velicogna, I., van den Broeke, M. R., Monaghan, A. & Lenaerts, J. Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level rise. *Geophys. Res. Lett.* **38,** L05503 (2011).
9. Ettema, J. *et al.* Higher surface mass balance of the Greenland ice sheet revealed by high-resolution climate modeling. *Geophys. Res. Lett.* **36,** L12501 (2009).
10. Lenaerts, J. T. M., van den Broeke, M. R., van de Berg, W. J., van Meijgaard, E. & Munneke, P. K. A new, high-resolution surface mass balance map of Antarctica (1979–2010) based on regional atmospheric climate modeling. *Geophys. Res. Lett.* **39,** L04501 (2012).
11. King, M. A. *et al.* Lower satellite-gravimetry estimates of Antarctic sea-level contribution. *Nature* **491,** 586–589 (2012).
12. Whitehouse, P. L., Bentley, M. J. & Le Brocq, A. M. A deglacial model for Antarctica: geological constraints and glaciological modelling as a basis for a new model of Antarctic glacial isostatic adjustment. *Quat. Sci. Rev.* **32,** 1–24 (2012).
13. Ivins, E. R. *et al.* Antarctic contribution to sea-level rise observed by GRACE with improved GIA correction. *J. Geophys. Res.* http://dx.doi.org/10.1002/jgrb.50208 (in the press).
14. Thomas, I. D. *et al.* Widespread low rates of Antarctic glacial isostatic adjustment revealed by GPS observations. *Geophys. Res. Lett.* **38,** L22302 (2011).
15. Whitehouse, P. L., Bentley, M. J., Milne, G. A., King, M. A. & Thomas, I. D. A new glacial isostatic adjustment model for Antarctica: calibrated and tested using observations of relative sea-level change and present-day uplift rates. *Geophys. J. Int.* **190,** 1464–1482 (2012).
    Demonstrates that new GIA models for Antarctica, which have been central to reconciling mass-balance estimates, greatly improve the fit between modelled and observed (GPS) uplift rates.
16. Sasgen, I. *et al.* Antarctic ice-mass balance 2002 to 2011: regional re-analysis of GRACE satellite gravimetry measurements with improved estimate of glacial-isostatic adjustment. *Cryosphere Discuss.* **6,** 3703–3732 (2012).
17. Horwath, M., Legresy, B., Remy, F., Blarel, F. & Lemoine, J. M. Consistent patterns of Antarctic ice sheet interannual variations from ENVISAT radar altimetry and GRACE satellite gravimetry. *Geophys. J. Int.* **189,** 863–876 (2012).
18. Zwally, H. J. & Giovinetto, M. B. Overview and assessment of Antarctic ice-sheet mass balance estimates: 1992–2009. *Surv. Geophys.* **32,** 351–376 (2011).
19. Zwally, H. J. *et al.* Mass balance of Antarctic ice sheet 1992 to 2008 from ERS and ICESat: gains exceed losses. *ISMASS 2012 Workshop* (2012); available at http://www.climate-cryosphere.org/en/events/2012/ISMASS/AntarcticIceSheet.html.
20. Sasgen, I. *et al.* Timing and origin of recent regional ice-mass loss in Greenland. *Earth Planet. Sci. Lett.* **333-334,** 293–303 (2012).
21. Ritz, C., Rommelaere, V. & Dumas, C. Modeling the evolution of Antarctic ice sheet over the last 420 000 years: implications for altitude changes in the Vostok region. *J. Geophys. Res.* **106,** 31943–31964 (2001).
22. Schoof, C. Ice sheet grounding line dynamics: steady states, stability, and hysteresis. *J. Geophys. Res.* **112,** F03S28 (2007).
23. Pollard, D. & Deconto, R. M. Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458,** 329–332 (2009).
24. Bueler, E. & Brown, J. Shallow shelf approximation as a "sliding law" in a thermomechanically coupled ice sheet model. *J. Geophys. Res.* **114,** F03008 (2009).
25. Pattyn, F. A new three-dimensional higher-order thermomechanical ice sheet model: basic sensitivity, ice stream development and ice flow across subglacial lakes. *J. Geophys. Res.* **108** (B8), 2382 (2003).
26. Blatter, H. Velocity and stress fields in grounded glaciers: a simple algorithm for including deviatoric stress gradients. *J. Glaciol.* **41,** 333–344 (1995).
27. Gillet-Chaulet, F. *et al.* Greenland Ice Sheet contribution to sea-level rise from a new-generation ice-sheet model. *Cryosphere* **6,** 1561–1576 (2012).

Represents the first complete implementation of full Stokes in dynamical ice-sheet models.
28. Larour, E., Seroussi, H., Morlighem, M. & Rignot, E. Continental scale, high order, high spatial resolution, ice sheet modeling using the Ice Sheet System Model (ISSM). *J. Geophys. Res.* **117,** F01022 (2012).
29. Cornford, S. L. *et al.* Adaptive mesh, finite volume modeling of marine ice sheets. *J. Comput. Phys.* **232,** 529–549 (2013).
A complete and correct implementation of 3D grounding line dynamics applied to Pine Island glacier for a loss of ice shelf buttressing, uniquely showing large grounding-line retreat.
30. Moon, T. & Joughin, I. Smith, B. & Howat, I. 21st-century evolution of Greenland outlet glacier velocities. *Science* **336,** 576–578 (2012).
31. Gogineni, P. CReSIS Data Products. http://data.cresis.ku.edu/ (2012).
32. Arthern, R. J. & Gudmundsson, G. H. Initialization of ice-sheet forecasts viewed as an inverse Robin problem. *J. Glaciol.* **56,** 527–533 (2010).
33. Scambos, T. A., Bohlander, J. A., Shuman, C. A. & Skvarca, P. Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophys. Res. Lett.* **31,** L18402 (2004).
34. Rignot, E. *et al.* Recent ice loss from the Fleming and other glaciers, Wordie Bay, West Antarctic Peninsula. *Geophys. Res. Lett.* **32,** L07502 (2005).
35. Jacobs, S. S., Jenkins, A., Giulivi, C. F. & Dutrieux, P. Stronger ocean circulation and increased melting under Pine Island Glacier ice shelf. *Nature Geosci.* **4,** 519–523 (2011).
36. MacAyeal, D. R. Scambos, T. A., Hulbe, C. L. & Fahnestock, M. A. Catastrophic ice-shelf break-up by an ice-shelf-fragment-capsize mechanism. *J. Glaciol.* **49,** 22–36 (2003).
37. Joughin, I., Smith, B. E. & Holland, D. M. Sensitivity of 21st century sea level to ocean-induced thinning of Pine Island Glacier, Antarctica. *Geophys. Res. Lett.* **37,** L20502 (2010).
38. Rignot, E. *et al.* Recent Antarctic ice mass loss from radar interferometry and regional climate modelling. *Nature Geosci.* **1,** 106–110 (2008).
39. Pattyn, F. *et al.* Grounding-line migration in plan-view marine ice-sheet models: results of the ice2sea MISMIP3d intercomparison. *J. Glaciol.* (in the press).
A community inter-comparison exercise that shows the capabilities of current ice-sheet models for robustly simulating grounding-line migration, which is key for predicting marine ice-sheet behaviour.
40. Greischar, L. L. & Bentley, C. R. Isostatic equilibrium grounding line between the West Antarctic inland ice-sheet and the Ross ice shelf. *Nature* **283,** 651–654 (1980).
41. Gomez, N., Mitrovica, J. X., Huybers, P. & Clark, P. U. Sea level as a stabilizing factor for marine-ice-sheet grounding lines. *Nature Geosci.* **3,** 850–853 (2010).
42. Gomez, N., Pollard, D., Mitrovica, J. X., Huybers, P. & Clark, P. U. Evolution of a coupled marine ice sheet-sea level model. *J. Geophys. Res.* **117,** F01013 (2012).
43. Das, S. B. *et al.* Fracture propagation to the base of the Greenland ice sheet during supraglacial lake drainage. *Science* **320,** 778–781 (2008).
44. Schoof, C. Ice sheet acceleration driven by melt supply variability. *Nature* **468,** 803–806 (2010).
Shows the important role of the ice sheet–ice shelf transition zone in controlling marine ice-sheet dynamics (in particular, stability/instability).
45. Sundal, A. *et al.* Melt-induced speed-up of Greenland ice sheet offset by efficient subglacial drainage. *Nature* **469,** 521–524 (2011).
46. Bartholomew, I. *et al.* Short-term variability in Greenland Ice Sheet motion forced by time-varying meltwater drainage: implications for the relationship between subglacial drainage system behavior and ice velocity. *J. Geophys. Res.* **117,** F03002 (2012).
47. Amundson, J. & Truffer, M. A unifying framework for iceberg-calving models. *J. Glaciol.* **56,** 822–830 (2010).
48. Hindmarsh, R. C. A. An observationally validated theory of viscous flow dynamics at the ice-shelf calving front. *J. Glaciol.* **58,** 375–387 (2012).
49. Bassis, J. N. The statistical physics of iceberg calving and the emergence of universal calving laws. *J. Glaciol.* **57,** 3–16 (2011).
50. Levermann, A. *et al.* Kinematic first-order calving law implies potential for abrupt ice-shelf retreat. *Cryosphere* **6,** 273–286 (2012).
51. Benn, D. I., Warren, C. R. & Mottram, R. H. Calving processes and the dynamics of calving glaciers. *Earth Sci. Rev.* **82,** 143–179 (2007).
52. Nick, F. M., Vieli, A., Howat, I. M. & Joughin, I. Large-scale changes in Greenland outlet glacier dynamics triggered at the terminus. *Nature Geosci.* **2,** 110–114 (2009).
53. Goelzer, H. *et al.* Millennial total sea-level commitments projected with the Earth system model of intermediate complexity LOVECLIM. *Environ. Res. Lett.* **7,** 045401 (2012).
54. Nghiem, S. V. *et al.* The extreme melt across the Greenland ice sheet in 2012. *Geophys. Res. Lett.* **39,** L20502 (2012).
Key paper documenting this large-scale Greenland melt event that was unprecedented in the modern satellite record.
55. Screen, J. A., Deser, C. & Simmonds, I. Local and remote controls on observed Arctic warming. *Geophys. Res. Lett.* **39,** L10709 (2011).
Provides strong observational and model evidence of symptoms and causes of the recent amplified Arctic warming.
56. Yoshimori, M. & Abe-Ouchi, A. Sources of spread in multimodel projections of the Greenland ice sheet surface mass balance. *J. Clim.* **25,** 1157–1175 (2012).
57. Harper, N., Humphrey, N. F., Pfeffer, W. T., Brown, J. & Fettweis, X. Greenland ice-sheet contribution to sea-level rise buffered by meltwater storage in firn. *Nature* **491,** 240–243 (2012).
58. Fettweis, X. *et al.* Estimating Greenland ice sheet surface mass balance contribution to future sea level rise using the regional atmospheric climate model MAR. *Cryosphere* **7,** 469–489 (2013).
59. Price, S. F., Payne, A. J., Howat, I. M. & Smith, B. E. Committed sea-level rise for the next century from Greenland ice sheet dynamics during the past decade. *Proc. Natl Acad. Sci. USA* **108,** 8978–8983 (2011).
60. Nick, F. M. *et al.* Future sea-level rise from Greenland's main outlet glaciers in a warming climate. *Nature* **497,** 235–238 (2013).
61. Goelzer, H. *et al.* Sensitivity of Greenland ice sheet projections to model formulations. *J. Glaciol.*. (in the press).
62. Bindschadler, R. A. *et al.* Ice sheet model sensitivities to environmental forcing and their use in projecting future sea level (the SeaRISE project). *J. Glaciol.* **59,** 195–224 (2013).
63. Arneborg, L., Wåhlin, A. K., Björk, G., Liljebladh, B. & Orsi, A. H. Persistent inflow of warm water onto the central Amundsen shelf. *Nature Geosci.* **5,** 876–880 (2012).
64. Bengtsson, L., Koumoutsaris, S. & Hodges, K. Large-scale surface mass balance of ice sheets from a comprehensive atmospheric model. *Surv. Geophys.* **32,** 459–474 (2011).
65. Winkelmann, R., Levermann, A., Martin, M. A. & Frieler, K. Increased future ice discharge from Antarctica owing to higher snowfall. *Nature* **492,** 239–242 (2012).
66. Little, C., Oppenheimer, M. & Urban, N. M. Upper bounds on twenty-first-century Antarctic ice loss assessed using a probabilistic framework. *Nature Clim. Change* http://dx.doi.org/10.1038/nclimate1845 (published online 17 March 2013).
67. Church, J. A. *et al.* Revisiting the Earth's sea-level and energy budgets from 1961 to 2008. *Geophys. Res. Lett.* **38,** L18601 (2011).
A good and recent (though the numbers are already outdated in many cases) review of all contributions to SLR.
68. Hock, R., de Woul, M., Radić, V. & Dyurgerov, M. Mountain glaciers and ice caps around Antarctica make a large sea-level rise contribution. *Geophys. Res. Lett.* **36,** L07501 (2009).
69. Dyurgerov, M. B. Reanalysis of glacier changes: from the IGY to the IPY, 1960–2008. *Data Glaciol. Studies* **108,** 5–116 (2010).
70. Cogley, J. G. Geodetic and direct mass-balance measurements: comparison and joint analysis. *Ann. Glaciol.* **50,** 96–100 (2009).
71. Jacob, T., Wahr, J., Pfeffer, W. T. & Swenson, S. Recent contributions of glaciers and ice caps to sea level rise. *Nature* **482,** 514–518 (2012).
72. Cogley, J. G. in *The Future of the World's Climate* 2nd edn (eds Henderson-Sellers, A. & McGuffie, K.) 197–222 (Elsevier, 2012).
73. Arendt, A. *et al. Randolph Glacier Inventory: A Dataset of Global Glacier Outlines* Version 2.0 (GLIMS Technical Report, Global Land Ice Measurements from Space, Boulder, 2012); available at http://www.glims.org/RGI/.
74. Gardner, A. S. *et al.* A reconciled estimate of glacier contributions to sea level rise: 2003 to 2009. *Science* **340,** 852–857 (2013).
Presents a consensus estimate of the contributions of glaciers and ice caps to sea-level rise that reconciles the disparate estimates previously available from the different techniques.
75. Meehl, G. A. *et al.* Relative outcomes of climate change mitigation related to global temperature versus sea level rise. *Nature Clim. Change* **2,** 576–580 (2012).
76. Gouretski, V. & Koltermann, K. P. How much is the ocean really warming? *Geophys. Res. Lett.* **34,** L01610 (2007).
77. Domingues, C. M. *et al.* Improved estimates of upper-ocean warming and multi-decadal sea-level rise. *Nature* **453,** 1090–1093 (2008).
78. von Schuckmann, K. & Le Traon, P.-Y. How well can we derive global ocean indicators from Argo data? *Ocean Sci. Discuss* **8,** 999–1024 (2011).
79. Leuliette, E. W. & Willis, J. K. Balancing the sea level budget. *Oceanography (Wash. DC)* **24,** 122–129 (2011).
80. Roemmich, D. & Gilson, J. The global ocean imprint of ENSO. *Geophys. Res. Lett.* **38,** L13606 (2011).
81. Wada, Y. *et al.* Past and future contribution of global groundwater depletion to sea-level rise. *Geophys. Res. Lett.* **39,** L09402 (2012).
82. Pokhrel, Y. N. *et al.* Model estimates of sea-level change due to anthropogenic impacts on terrestrial water storage. *Nature Geosci.* **5,** 389–392 (2012).
83. Konikow, L. F. Overestimated water storage. *Nature Geosci.* **6,** 3–4 (2012).
84. Remy, F., Flament, T., Blarel, F. & Benveniste, J. Radar altimetry measurements over Antarctic ice sheet: a focus on antenna polarization and change in backscatter problems. *Adv. Space Res.* **50,** 998–1006 (2012).
85. Nicholls, R. J. *et al.* Sea-level rise and its possible impacts given a 'beyond 4°C world' in the twenty-first century. *Proc. R. Soc. Lond. A* **369,** 161–181 (2011).
86. Joughin, I., Smith, B., Howat, I., Scambos, T. & Moon, T. Greenland flow variability from ice-sheet-wide velocity mapping. *J. Glaciol.* **56,** 415–430 (2010).
87. Purkey, S. G. & Johnson, G. C. Warming of global abyssal and deep Southern Ocean waters between the 1990s and 2000s: contributions to global heat and sea level rise budgets. *J. Clim.* **23,** 6336–6351 (2010).
88. Levitus, S. *et al.* World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophys. Res. Lett.* **39,** L10603 (2012).
89. Harig, C. & Simons, F. J. Mapping Greenland's mass loss in space and time. *Proc. Natl Acad. Sci. USA* **109,** 19934–19937 (2012).
90. Ewert, H., Groh, A. & Dietrich, R. Volume and mass changes of the Greenland ice sheet inferred from ICESat and GRACE. *J. Geodyn.* **59–60,** 111–123 (2012).
91. Moss, R. H. *et al.* The next generation of scenarios for climate change research and assessment. *Nature* **463,** 747–756 (2010).
92. Farrell, W. E. & Clark, J. A. On postglacial sea level. *Geophys. J. R. Astron. Soc.* **46,** 647–667 (1976).
93. Kendall, R. A., Mitrovica, J. X. & Milne, G. A. On post-glacial sea level — II. Numerical formulation and comparative results on spherically symmetric models. *Geophys. J. Int.* **161,** 679–706 (2005).
94. Wahr, J., Wingham, D. & Bentley, C. A method of combining ICESat and GRACE satellite data to constrain Antarctic mass balance. *J. Geophys. Res.* **105,** 16279–16294 (2000).

95. Simpson, M. J. R., Wake, L., Milne, G. A. & Huybrechts, P. The influence of decadal- to millennial-scale ice mass changes on present-day vertical land motion in Greenland: Implications for the interpretation of GPS observations. *J. Geophys. Res.* **116,** B02406 (2011).
96. Dietrich, R. *et al.* Rapid crustal uplift in Patagonia due to enhanced ice loss. *Earth Planet. Sci. Lett.* **289,** 22–29 (2010).
97. Sato, T. *et al.* Reevaluation of the viscoelastic and elastic responses to the past and present-day ice changes in Southeast Alaska. *Tectonophysics* **511,** 79–88 (2011).
98. Morelli, A. & Danesi, S. Seismological imaging of the Antarctic continental lithosphere: a review. *Global Planet. Change* **42,** 155–165 (2004).
99. Tarasov, L., Dyke, A. S., Neal, R. M. & Peltier, W. R. A data-calibrated distribution of deglacial chronologies for the North American ice complex from glaciological modeling. *Earth Planet. Sci. Lett.* **315–316,** 30–40 (2012).
100. Gudmundsson, G. H. *et al.* The stability of grounding lines on retrograde slopes. *Cryosphere* **6,** 1497–1505 (2012).

Acknowledgements The work presented here is based on the Ice-Sheet Mass Balance and Sea Level (ISMASS) workshop that was held in Portland, Oregon, USA, on 14 July 2012. This workshop was jointly organized by the Scientific Committee on Antarctic Research (SCAR), the International Arctic Science Committee (IASC) and the Word Climate Research Programme (WCRP), and was co-sponsored by the International Council for Science (ICSU), SCAR, IASC, WCRP, the International Glaciological Society (IGS) and the International Association of Cryospheric Sciences (IACS), with support from Climate and Cryosphere (CliC) and the Association of Polar Early Career Scientists (APECS).

Author Contributions E.H. coordinated the study, E.H., F.J.N. and F.P. led the writing, and all authors contributed to the writing and discussion of ideas.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.H. (ehanna@sheffield.ac.uk).

ARTICLE

doi:10.1038/nature12200

The oldest known primate skeleton and early haplorhine evolution

Xijun Ni[1,2], Daniel L. Gebo[3], Marian Dagosto[4], Jin Meng[2], Paul Tafforeau[5], John J. Flynn[2] & K. Christopher Beard[6]

Reconstructing the earliest phases of primate evolution has been impeded by gaps in the fossil record, so that disagreements persist regarding the palaeobiology and phylogenetic relationships of the earliest primates. Here we report the discovery of a nearly complete and partly articulated skeleton of a primitive haplorhine primate from the early Eocene of China, about 55 million years ago, the oldest fossil primate of this quality ever recovered. Coupled with detailed morphological examination using propagation phase contrast X-ray synchrotron microtomography, our phylogenetic analysis based on total available evidence indicates that this fossil is the most basal known member of the tarsiiform clade. In addition to providing further support for an early dichotomy between the strepsirrhine and haplorhine clades, this new primate further constrains the age of divergence between tarsiiforms and anthropoids. It also strengthens the hypothesis that the earliest primates were probably diurnal, arboreal and primarily insectivorous mammals the size of modern pygmy mouse lemurs.

Primates Linnaeus, 1758
Haplorhini Pocock, 1918
Tarsiiformes Gregory, 1915
Archicebidae fam. nov.
Archicebus achilles gen. et sp. nov.

Etymology. Generic name is derived from arche, Greek for beginning, and cebus, new Latin from Greek, for long-tailed monkey. Specific epithet is from Achilles, in allusion to the very interesting anthropoid-like heel bone (calcaneus) of the type species.

Holotype. IVPP V18618, a partial skeleton preserved as part and counterpart (Fig. 1 and Supplementary Information).

Locality and horizon. The lower part of the lower Eocene Yangxi Formation in Jingzhou area, Hubei Province, China. Bumbanian Asian Land Mammal Age, 55.8–54.8 million years (Myr) ago[1].

Diagnosis. Small haplorhine primate with rounded braincase; short snout; vertically implanted upper canine (C^1); four premolars in each jaw quadrant; long hindlimbs; long feet (especially the metatarsus); and a long tail. Among other basal primates, differs from *Donrussellia*, *Marcgodinotius* and *Asiadapis* in having a single-rooted lower second premolar (P_2), and differs from *Teilhardina belgica*, '*Teilhardina*' *americana* and '*Teilhardina*' *brandti* in having a less-reduced P_1. Further differs from *T. belgica* in having relatively shorter and broader distal calcaneus and smaller peroneal tubercle on the first metatarsal. Further differs from '*T.*' *americana* and '*T.*' *brandti* in having weaker cingulum and cingulid on upper and lower molars and lacking the *Nannopithex*-fold. Differs from *Teilhardina asiatica* in having a weaker P_4 metaconid, lower-crowned P_{3-4}, and a more prominent, hook-like mandibular angular process. Differs from *Teilhardina magnoliana* in having stronger mesial and distal cingula on the upper molars and a shorter talonid on P_4.

Description

This new early Eocene primate is a very small animal, with slender limbs and a long tail. The trunk is about 71 mm, the tail is more than 130 mm, and the skull is approximately 25 mm long and 17 mm wide.

Skull

The general shape of the skull is similar to that of *Teilhardina asiatica* and *Tetonius homunculus* (Fig. 2). A postorbital bar is present, but no postorbital septum. Relative to skull length, orbital diameter (7 mm) resembles that of *T. asiatica*[2], being proportionally smaller than those of most other tarsiiforms, and falling within the range of variation exhibited by extant diurnal primates[3] (Supplementary Information). As in other primates, the orbits are significantly convergent. The nasal fossa shows substantial reduction relative to the condition in outgroups. Preorbital snout length (4.7 mm) is short, as in *Tarsius*, *Tetonius*, *Shoshonius* and most anthropoids; in contrast, *Omomys*, *Necrolemur* and most strepsirrhines retain proportionally longer snouts. The left and right upper dental arcades are gently divergent, resembling those of *T. asiatica* and *Rooneyia*. *Tarsius*, *Necrolemur*, *Shoshonius* and other more anatomically derived tarsiiforms have bell-shaped palates, due to the combination of orbital hypertrophy and snout reduction[4]. The dentary is gracile, with a shallow, procumbent and unfused symphysis. The gonial part of the dentary bears a long, hook-like angular process with a very strong pterygoid crest on its medial side.

Dentition

The dentition of *A. achilles* shows a very primitive morphology (Fig. 2), being comparable to that of other phylogenetically basal primates such as *Teilhardina*, *Donrussellia*, *Marcgodinotius* and *Asiadapis*. An isolated lower central incisor, bearing a mesiodistally compressed root and labiolingually compressed (spatulate) and symmetrical crown, is associated with this specimen. C^1 has a vertically implanted root. Its crown projects well below the occlusal plane of the molars. C_1 is not preserved, but its alveolus indicates that this tooth is unreduced. As in other basal primates, there are four premolars in each upper and lower jaw quadrant. P^{1-2} are small, single-cusped and single-rooted teeth. The alveoli for P_{1-2} suggest that they are small, single-rooted and probably as simple as their upper counterparts. P^2 and P_2 are present in basal tarsiiforms and anthropoids, in which they

[1]Key Laboratory of Vertebrate Evolution and Human Origin, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, 142 Xi Zhi Men Wai Street, Beijing 100044, China. [2]Division of Paleontology and Richard Gilder Graduate School, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024, USA. [3]Department of Anthropology, Northern Illinois University, DeKalb, Illinois 60115, USA. [4]Department of Cell and Molecular Biology, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. [5]European Synchrotron Radiation Facility, 38043 Grenoble, France. [6]Section of Vertebrate Paleontology, Carnegie Museum of Natural History, 4400 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA.

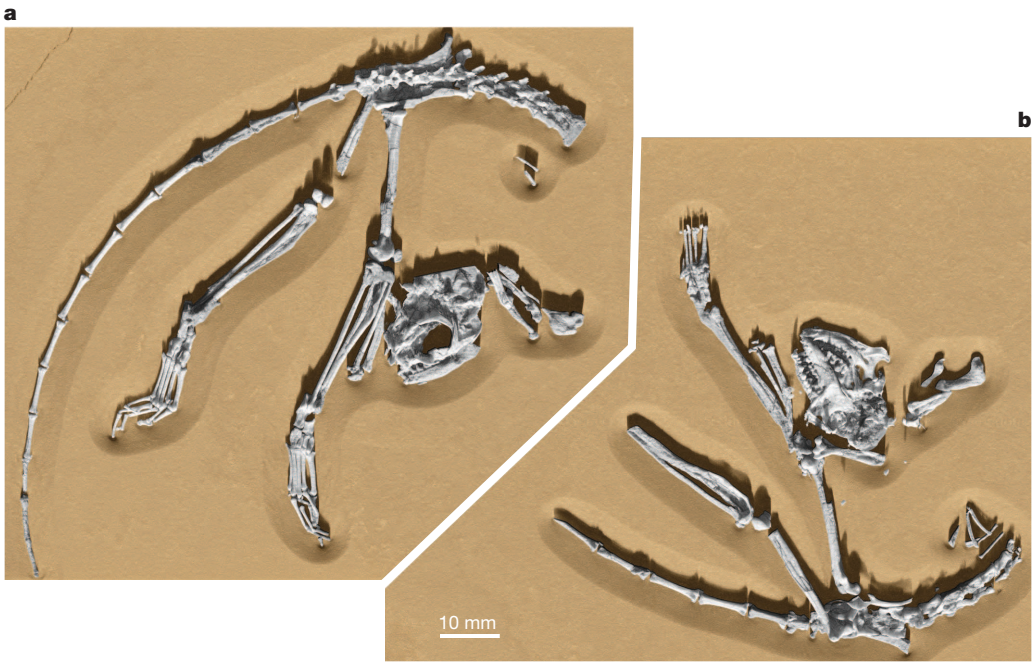


Figure 1 | Three-dimensional reconstruction of the type specimen (IVPP V18618) of *Archicebus achilles*. **a**, Slab-a, dorsal view of the skull, lumbar region and pelvis, laterodorsal view of the tail, posterior view of the left thigh, medial view of the left leg, plantar view of the left foot, lateral view of the right thigh, lateral view of the right leg, and dorsal view of the right foot. **b**, Slab-b, ventral view of the skull, lumbar region and pelvis, anterolateral view of the left thigh, and posteromedial view of the right thigh. Fossil bones are shown in light grey. Digital casts reconstructed from the preserved impressions are shown in darker grey than the actual bones. The bones yielding the impressions are either preserved on the counterpart or were lost during collection and/or preparation of the specimen.

are small, simple and usually single rooted[2,5–8]. In contrast, these teeth in haplorhine outgroups such as basal adapiforms are double rooted and only slightly smaller than P^3 and P_3 (refs 9–11). $P^{3–4}$ of *A. achilles* resemble those of *Teilhardina*. The paracones of these teeth are high and sharp, with rounded mesial borders and well-developed distal crests. The protocones of $P^{3–4}$ are large, but mesiodistally shorter than

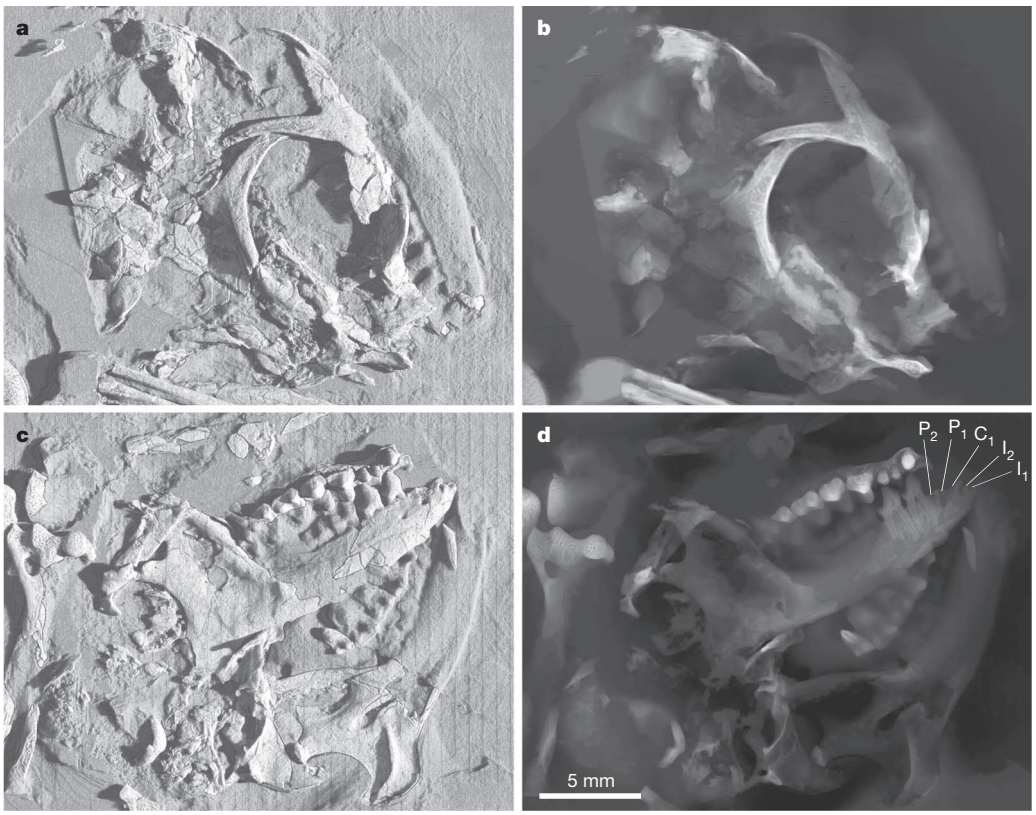


Figure 2 | The head region of *Archicebus achilles*. **a**, Dorsal view of the skull. **b**, Pseudo-radiograph rendering of the dorsal view of the skull. **c**, Ventral view of the skull, lingual view of the left mandible and lateral view of right mandible. **d**, Pseudo-radiograph rendering of the ventral view of the skull, lingual view of the left mandible and lateral view of right mandible.

those of *T. asiatica* and *T. belgica*. The trigonids of P_{3-4} bear a single major cusp, and their talonids are short and heel-like. The paraconids of P_{3-4} are low and weak. The metaconid is absent in P_3 and only weakly developed in P_4, as in *T. magnoliana* and '*Teilhardina*' *brandti*[5,12]. The upper molars closely resemble those of *Teilhardina* and other basal tarsiiforms. The distal borders of M^{1-2} are concave, their conules are large, and M^3 is small relative to M^{1-2}. M^{1-2} lack any development of a postprotocingulum or *Nannopithex*-fold, in contrast to the condition in *Tetonius*, '*T.*' *americana* and most other anaptomorphine omomyids, being more similar to *T. asiatica* and *T. belgica* in this regard.

Trunk and tail

Only the lumbar region of the trunk is well preserved, showing at least six or possibly seven lumbar vertebrae. The caudal region preserves 18 vertebrae. Considering the gradual reduction of caudal vertebral lengths distally, this new primate may have had over 30 caudal vertebrae in life, making its tail exceedingly long relative to head and trunk length.

Forelimb

The scapula possesses a glenoid fossa that is tear-shaped, and a long coracoid process that almost exceeds the craniocaudal length of the glenoid fossa. The humerus (15.7 mm) is relatively short. The humeral head is oval, projecting slightly above the lesser and greater tuberosities, which are separated by a broad and shallow bicipital groove. This morphology resembles other Palaeogene primates[13–16]. The elbow joint bears a rounded capitulum clearly distinguished from the trochlear joint surface by the zona conoidea, a key feature of primates. A significant capitular tail extends laterally from the capitulum. The trochlea is oblique with a 'downturned' medial rim, a typical feature among early haplorhine primates[14,17]. In distal view, the distal articular surface is 'waisted', as in anthropoids[16–18]. A dorsoepitrochlear fossa is present. This feature is absent in tarsiers but shared by omomyiforms and most basal anthropoids[16,19]. The entepicondylar foramen is large and located above the medial part of the trochlea, a primitive condition for primates.

The ulna has a short olecranon process, a strong and straight shaft, and a narrow distal end, whereas the radial head has a wide articular circumference and a long and angled radial neck. Relative to the humerus, radial length (estimated as 16.0 mm, brachial index approximately 102) is unremarkable, except to demonstrate that *A. achilles* does not exhibit the relatively long radius of *Tarsius*. Only impressions of the carpals are preserved. As in other haplorhines, the hamate has a mediolaterally oriented spiral facet for the triquetrum.

Hindlimb

The hindlimbs are almost completely preserved (Figs 1, 3). The ilia are long and narrow, with slightly concave gluteal surfaces. This pelvic shape resembles omomyids, tarsiers and some extant strepsirrhines, but differs from the broader bladed ilia of anthropoids, adapiforms, lemurids and indriids[16,20–23]. The two iliac crests are slightly divergent and extend cranially over the sacral wing for a short distance. A large inferior iliac spine lies on the cranial side of the acetabulum. The ischium is long, straight and stout, being caudally directed with no significant dorsal projection.

The thigh and lower leg (femur, 27.0 mm; tibia, 30.1 mm; fibula, 29.1 mm) are very long relative to the arm and forearm. The intermembral index (IMI; 55) is equivalent to that of the most specialized extant vertical clinging and leaping primates (for example, tarsiers and galagos), and lower than estimated for *Shoshonius*[14].

The femur is slender, and the femoral head is semi-cylindrical with the proximal articular surface extending onto the femoral neck as in omomyids and extant vertical clinging and leaping primates[13,14,22]. A large fovea capitis femoris occurs on the medial side of the femoral head. The femoral neck is moderately long (3.7 mm) and forms an angle of 49.7° relative to the femoral shaft. The broad greater trochanter extends above the femoral head. Its lateral border is thick, flaring laterally and being confluent with a triangular-shaped third trochanter distally. The trochanteric fossa is moderately long. The lesser trochanter extends posteromedially and forms an angle of 40.6° relative to the femoral shaft. The proximal part of the femoral shaft is not as anteriorly bowed as in omomyids or microchoerids, resembling anthropoids instead. Distal to the third trochanter, the femoral shaft is straight and robust. The knee is quite tall, with an elevated lateral patellar rim and a long and broad patellar articular facet, additional similarities to frequently leaping primates.

The tibia is quite straight and lacks the marked S-shaped curvature observed in *Shoshonius*[14]. The cnemial crest is strong and long, extending distally over half of the total tibial length. The intercondylar eminence of the tibial plateau has two spines, a haplorhine feature[24]. The crural index (109–113) closely resembles those of leaper-quadrupeds

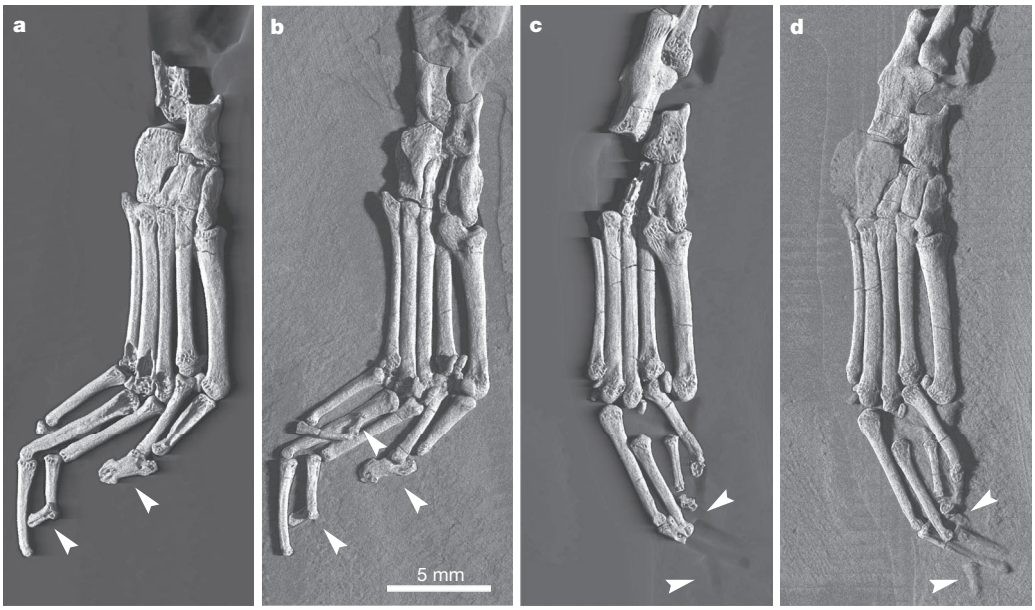


Figure 3 | The foot region of *Archicebus achilles*. **a**, Dorsal view of the left foot (reversed). **b**, Plantar view of the left foot. **c**, Plantar view of the right foot (reversed). **d**, Dorsal view of the right foot. Arrowheads indicate the scutiform distal phalanges of the big, second, third and fifth toes.

such as *Galagoides* and *Microcebus*, but is higher than that of small vertical clinging and leaping primates[14]. The fibula is straight, robust and closely apposed to the tibia for ~36% of the length of the distal shaft. The tibia and fibula are unfused and lack any prominent tibio-fibular scar, in contrast to omomyids. The distal tibia and fibula exhibit a standard haplorhine tibio-fibular mortise, the likely primitive condition for primates. The tibial malleolus is shortened and slightly angled posteriorly, being similar to haplorhines, and in sharp contrast to strepsirrhine primates[25].

The foot (Fig. 3, estimated at 33.5 mm long, from the calcaneal tuber to the tip of the fourth digit) is 36.6% of total hindlimb length, a similarity to primates with particularly long metatarsals (for example, callitrichid platyrrhines) but also to primates with a long tarsal region (for example, galagos and tarsiers)[26]. Relative to body mass, *A. achilles* has a moderately short tarsus, closely similar to those of extant anthropoids; a very long metatarsus, comparable to anthropoids and tupaiids (and different from lemuriforms, adapiforms, or *Tarsius*); and a long phalangeal region, most similar to *Tarsius* (Supplementary Information). This combination of foot proportions is unique among living and fossil primates and their nearest relatives.

The width to length ratio of the calcaneus is 40.2%, a value very close to eosimiids, but higher than other tarsiiforms and lower than platyrrhines and strepsirrhines[14,22,27–30]. The middle and distal parts of the calcaneus are proportionally wide, the width to length ratio of the posterior facet is high, and the heel is proportionally short relative to posterior calcaneal facet length. These features are very similar to eosimiids and platyrrhines, but differ from other tarsiiforms and strepsirrhines. The distal region of the calcaneus is moderately elongated (52.0% of total calcaneal length), falling in the range of tarsiiforms and eosimiids[28]. The calcaneocuboid joint is fan-shaped, a primitive primate condition[30]. The talus is not preserved, but its impression indicates that this bone had a broad head, a long neck and a moderately developed posterior trochlear shelf. These features are present in most tarsiiforms and eosimiids[14,22,28–30]. The moderately elongated cuboid, navicular and entocuneiform are similar in morphology to those of other tarsiiforms, being only slightly elongated. The navicular-cuboid facet contacts only the ectocuneiform facet of the navicular, a haplorhine characteristic[13,17,31].

The first metatarsal–entocuneiform joint is curved, with a narrow joint arc, a similarity shared with other tarsiiforms. However, the curvature of this joint surface is slightly asymmetrical, a similarity to a specimen tentatively referred to an eosimiid or a tarsier[32]. The peroneal tubercle is moderately long, high and wide, being similar to adapiforms, microchoerids and eosimiids, in contrast to the narrow or pointed peroneal tubercle observed in omomyids or the wide proximal ends of platyrrhines.

The proximal and middle pedal phalanges are long and fairly straight in lateral view. They lack the greater curvature of most primate phalanges. The fourth digit is the longest (ectaxony), a similarity to tarsiers and lemuriforms, in contrast to third digit elongation (mesaxony) that characterizes adapiforms and anthropoids. The distal phalanges of the first, third and fifth digits, and a fine impression of the second, are preserved (Fig. 3 and Supplementary Information). All of these distal phalanges are scutiform with dorsoplantarly compressed and expanded distal apical tufts, indicating the presence of flat nails.

Phylogeny

Archicebus achilles possesses a unique mosaic of haplorhine features, some of which resemble anthropoids whereas others resemble tarsiiform primates. For example, in terms of calcaneal shape and metatarsal proportions within the foot, the new taxon recalls anthropoid primates, whereas its skull, dentition and many aspects of its appendicular skeleton resemble tarsiiforms. This combination of anthropoid-like and tarsiiform-like features in a single taxon is unique and unexpected, posing novel possibilities for reconstructing how modern tarsiers and anthropoids evolved their diagnostic characters.

Detailed comparative anatomical research and revised phylogenetic analysis based on an updated, large, combined morphological and molecular character data matrix place *A. achilles* as the most basal member of a monophyletic tarsiiform haplorhine clade (Fig. 4 and Supplementary Information). Anthropoidea is monophyletic and sister group to the tarsiiforms. *Archicebus achilles* therefore helps constrain and push back the age of the split between tarsiiforms and anthropoids, and an early division between strepsirrhine and haplorhine primates[2,33–35] is supported. Furthermore, adapiforms are more closely related to lemuriforms than to anthropoids or any other haplorhines.

Plesiadapiforms, traditionally regarded as archaic primates[9,36], are not even stem primates, corroborating the now common practice of excluding plesiadapiforms from the order Primates[37–39].

Adaptive profile

Archicebus achilles (estimated body mass ~20–30 g, Supplementary Information) is as small as the modern pygmy mouse lemur[40]. Its large canines and sharply pointed premolars with well-developed shearing crests suggest a primarily insectivorous diet. The moderately large and convergent orbits of *A. achilles* indicate that the visual system had an important role during ingestion and locomotion, as is the case in modern primates. However, the absence of any marked orbital hypertrophy, which occurs uniformly in extant nocturnal haplorhines, indicates a diurnal activity pattern for *A. achilles* (Supplementary Information). Diurnality has also been suggested for *T. asiatica*, another basal haplorhine primate from Asia[2].

The postcranium of *A. achilles* shows many hindlimb features associated with frequent leaping, such as a long leg, a semi-cylindrical femoral head with a stout and less oblique femoral neck, a tall knee, and a closely apposed fibula. However, the long coronoid process of

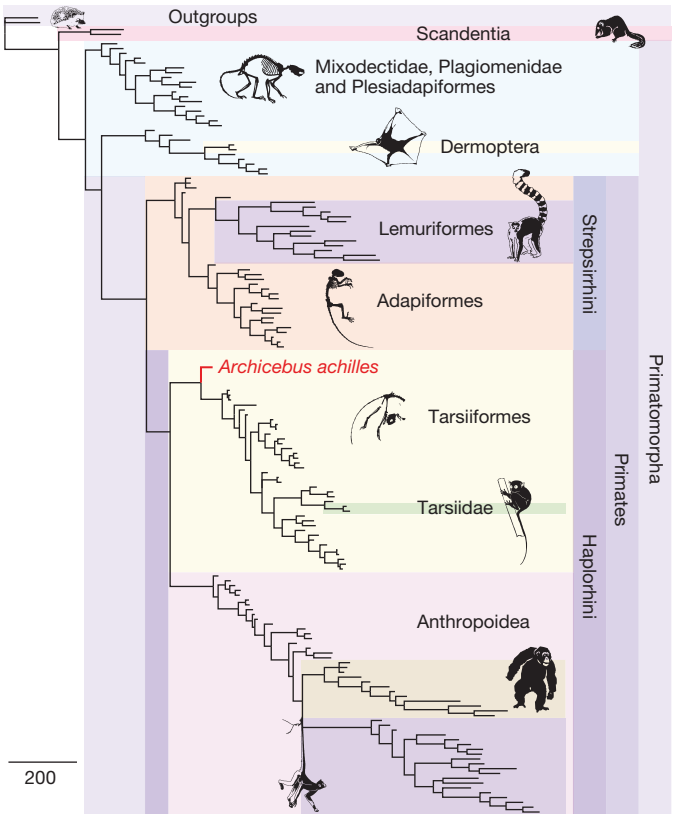


Figure 4 | Summary phylogeny of 157 mammals. Parsimony analysis is based on a data matrix including 1,186 morphological characters and 658 molecular characters of long and short interspersed nuclear elements scored for 119 fossil and 38 living taxa. Topology of extant treeshrews, flying lemurs and primates based on gene supermatrix is used as backbone constraint (Supplementary Information). Scale bar, 200 characters.

the scapula, a moderately rounded humeral head, a long and straight ischium, a high crural index, and the long metatarsal and phalangeal proportions of the foot of *A. achilles* are all linked to more generalized arboreal quadrupedal locomotion (or grasp-leaping), in contrast to the morphology of specialized vertical clinging and leaping primates such as galagids and tarsiers[14,19,29,41,42].

A long-standing idea holds that basal members of the major primate radiations are likely to be morphologically very similar to each other[9,43]. From this perspective, our reconstructed adaptive profile of the remarkably complete and well preserved skeleton of *A. achilles* may well mirror that of other phylogenetically basal primates, including the most basal anthropoids, the most basal haplorhines, and even the last common ancestor of all primates.

Received 1 February; accepted 18 April 2013.

1. Wang, Y. *et al.* Early Paleogene stratigraphic sequences, mammalian evolution and its response to environmental changes in Erlian Basin, Inner Mongolia, China. *Sci. China Earth Sci.* **53,** 1918–1926 (2010).
2. Ni, X., Wang, Y., Hu, Y. & Li, C. A euprimate skull from the early Eocene of China. *Nature* **427,** 65–68 (2004).
3. Kay, R. F. & Kirk, E. C. Osteological evidence for the evolution of activity pattern and visual acuity in primates. *Am. J. Phys. Anthropol.* **113,** 235–262 (2000).
4. Rossie, J. B., Ni, X. & Beard, K. C. Cranial remains of an Eocene tarsier. *Proc. Natl Acad. Sci. USA* **103,** 4381–4385 (2006).
5. Rose, K. D., Chester, S. G. B., Dunn, R. H., Boyer, D. M. & Bloch, J. I. New fossils of the oldest North American euprimate *Teilhardina brandti* (Omomyidae) from the Paleocene–Eocene thermal maximum. *Am. J. Phys. Anthropol.* **146,** 281–305 (2011).
6. Gingerich, P. D. Dental variation in early Eocene *Teilhardinal belgica*, with notes on the anterior dentition of some early tarsiiformes. *Folia Primatol.* **28,** 144–153 (1977).
7. Beard, K. C. & Wang, J. The eosimiid primates (Anthropoidea) of the Heti Formation, Yuanqu Basin, Shanxi and Henan Provinces, People's Republic of China. *J. Hum. Evol.* **46,** 401–432 (2004).
8. Beard, K. C., Tong, Y., Dawson, M. R., Wang, J. & Huang, X. Earliest complete dentition of an anthropoid primate from the late middle Eocene of Shanxi Province, China. *Science* **272,** 82–85 (1996).
9. Szalay, F. S. & Delson, E. *Evolutionary History of the Primates* (Academic, 1979).
10. Godinot, M. A summary of adapiform systematics and phylogeny. *Folia Primatol.* **69,** 218–249 (1998).
11. Rose, K. D. *et al.* Early Eocene primates from Gujarat, India. *J. Hum. Evol.* **56,** 366–404 (2009).
12. Beard, K. C. The oldest North American primate and mammalian biogeography during the Paleocene-Eocene Thermal Maximum. *Proc. Natl Acad. Sci. USA* **105,** 3815–3818 (2008).
13. Dagosto, M. in *Postcranial Adaptation in Nonhuman Primates* (ed. Gebo, D. L.) 150–174 (Northern Illinois Univ. Press, 1993).
14. Dagosto, M., Gebo, D. L. & Beard, K. C. Revision of the Wind River faunas, early Eocene of central Wyoming. Part 14. Postcranium of *Shoshonius cooperi* (Mammalia, Primates). *Annals Carnegie Museum* **68,** 175–211 (1999).
15. Fleagle, J. G. & Simons, E. L. The humerus of *Aegyptopithecus zeuxis*: a primitive anthropoid. *Am. J. Phys. Anthropol.* **59,** 175–193 (1982).
16. Fleagle, J. G. & Simons, E. L. Limb skeleton and locomotor adaptations of *Apidium phiomense*, an Oligocene anthropoid from Egypt. *Am. J. Phys. Anthropol.* **97,** 235–289 (1995).
17. Dagosto, M. & Gebo, D. L. in *Anthropoid Origins* (eds Fleagle, J. G. & Kay, R. K.) 567–593 (Plenum, 1994).
18. Gebo, D. L., Simons, E. L., Rasmussen, D. T. & Dagosto, M. in *Anthropoid Origins* (eds Fleagle, J. G. & Kay, R. F.) 203–233 (Plenum, 1994).
19. Szalay, F. S. & Dagosto, M. Locomotor adaptations as reflected on the humerus of Paleogene primates. *Folia Primatol.* **34,** 1–45 (1980).
20. Gregory, W. K. On the structure and relations of *Notharctus*, an American Eocene primate. *Memoirs of the American Museum of Natural History* series 3, 49–243 (1920).
21. Rose, K. D. & Walker, A. The skeleton of early Eocene *Cantius*, oldest lemuriform primate. *Am. J. Phys. Anthropol.* **66,** 73–89 (1985).
22. Anemone, R. L. & Covert, H. H. New skeletal remains of *Omomys* (Primates, Omomyidae): functional morphology of the hindlimb and locomotor behavior of a middle Eocene primate. *J. Hum. Evol.* **38,** 607–633 (2000).
23. Gebo, D. L., Dagosto, M., Beard, K. C. & Ni, X. New primate hind limb elements from the middle Eocene of China. *J. Hum. Evol.* **55,** 999–1014 (2008).
24. White, J. L. & Gebo, D. L. Unique proximal tibial morphology in strepsirrhine primates. *Am. J. Primatol.* **64,** 293–308 (2004).
25. Dagosto, M., Gebo, D. L., Ni, X., Qi, T. & Beard, K. C. in *Mammalian Evolutionary Morphology: A Tribute to Frederick S. Szalay* (eds Sargis, E. J. & Dagosto, M.) 315–324 (Springer, 2008).
26. Jouffroy, F. K. & Lessertisseur, J. in *Environment, Behavior, and Morphology: Dynamic Interactions in Primates* (eds Morbeck, M. E., Preuschoft, H. & Gomberg, N.) 143–181 (Gustav Fisher, 1979).
27. Covert, H. H. & Hamrick, M. W. Description of new skeletal remains of the early Eocene anaptomorphine primate *Absarokius* (Omomyidae) and a discussion about its adaptive profile. *J. Hum. Evol.* **25,** 351–362 (1993).
28. Gebo, D. L., Dagosto, M., Beard, K. C. & Qi, T. Middle Eocene primate tarsals from China: implications for Haplorhine evolution. *Am. J. Phys. Anthropol.* **116,** 83–107 (2001).
29. Gebo, D. L., Smith, T. & Dagosto, M. New postcranial elements for the earliest Eocene fossil primate *Teilhardina belgica*. *J. Hum. Evol.* **63,** 205–218 (2012).
30. Gebo, D. L., Dagosto, M., Beard, K. C., Qi, T. & Wang, J. The oldest known anthropoid postcranial fossils and the early evolution of higher primates. *Nature* **404,** 276–278 (2000).
31. Dagosto, M. Implications of postcranial evidence for the origin of euprimates. *J. Hum. Evol.* **17,** 35–56 (1988).
32. Gebo, D. L., Dagosto, M., Beard, K. C., Ni, X. & Qi, T. in *Elwyn Simons: A Search for Origins Developments in Primatology: Progress and Prospects* (eds Fleagle, J. G. & Gilbert, C. C.) 229–242 (Springer, 2008).
33. Kay, R. F., Ross, C. F. & Williams, B. A. Anthropoid origins. *Science* **275,** 797–804 (1997).
34. Ni, X. *et al.* A new tarkadectine primate from the Eocene of Inner Mongolia, China: phylogenetic and biogeographic implications. *Proc. R. Soc. B* **277,** 247–256 (2010).
35. Seiffert, E. R., Perry, J. M. G., Simons, E. L. & Boyer, D. M. Convergent evolution of anthropoid-like adaptations in Eocene adapiform primates. *Nature* **461,** 1118–1121 (2009).
36. Simons, E. L. *Primate Evolution: an Introduction to Man's Place in Nature* (Macmillan, 1972).
37. Fleagle, J. G. *Primate Adaptation and Evolution* 2nd edn, 1–596 (Academic, 1999).
38. Covert, H. H. in *The Primate Fossil Record* (ed. Hartwig, W. C.) 13–20 (Cambridge Univ. Press, 2002).
39. Beard, K. C. in *Mammal Phylogeny* Vol. 2 *Placentals* (eds Szalay, F. S., Novacek, M. J. & McKenna, M. C.) 129–150 (Springer, 1993).
40. Rasoloarison, R., Goodman, S. & Ganzhorn, J. Taxonomic revision of mouse lemurs (*Microcebus*) in the western portions of Madagascar. *Int. J. Primatol.* **21,** 963–1019 (2000).
41. Gebo, D. L. Locomotor diversity in prosimian primates. *Am. J. Primatol.* **13,** 271–281 (1987).
42. Gebo, D. L., Dagosto, M., Ni, X. & Beard, K. C. Species diversity and postcranial anatomy of Eocene primates from Shanghuang, China. *Evol. Anthropol.* **21,** 224–238 (2012).
43. Rose, K. D. The earliest primates. *Evol. Anthropol.* **3,** 159–173 (1994).

Supplementary Information is available in the online version of the paper.

Acknowledgements This project has been supported by the Strategic Priority Research Program of Chinese Academy of Sciences (CAS, XDB03020501), the National Basic Research Program of China (2012CB821904), the CAS 100-talent Program, the National Natural Science Foundation of China (40672009, 40872032), the US National Science Foundation (BCS 0820602), the ESRF (proposal ec347), and the Postdoctoral Research Fellowship Program of the American Museum of Natural History (AMNH). We are grateful to C. Li, Y. Wang, E. Delson, A. L. Rosenberger, E. Seiffert, M. T. Silcox and J. I. Bloch for helpful discussions. We thank C. Li and Q. Li for their assistance in the field, and C. Nemoz, T. Brochard and all the ID17 beamline team for their help during the synchrotron experiment. We thank the staff of the following museums for access to specimens: AMNH, Field Museum of Natural History, Chicago; Smithsonian Institution, Washington, D.C.; Carnegie Museum of Natural History, Pittsburgh; Royal Belgium Institute of Natural Sciences, Brussels.

Author Contributions X.N. designed the study, analysed the data and wrote the paper. K.C.B., J.M., D.L.G. and M.D. contributed extensively and equally to the work presented in this paper. P.T. performed synchrotron microtomography experiments and edited the manuscript. J.J.F. collected part of the data and edited the manuscript.

Author Information ZooBank accessions: urn:lsid:zoobank.org:act:884CBACC-B602-471A-A7B0-E2AF092BA6F8 (Archicebidae fam. nov.); urn:lsid:zoobank.org:act:163DE8EB-D691-49E4-A211-8ECF117756BD (*Archicebus* gen. nov.); urn:lsid:zoobank.org:act:105EE748-38DE-4709-A7D0-44FE0E3E2813 (*Archicebus achilles* sp. nov.). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to X.N. (nixijun@ivpp.ac.cn).

ARTICLE

doi:10.1038/nature12153

Rats maintain an overhead binocular field at the expense of constant fusion

Damian J. Wallace[1]*, David S. Greenberg[1]*, Juergen Sawinski[1]*, Stefanie Rulla[1], Giuseppe Notaro[1,2] & Jason N. D. Kerr[1,2]

Fusing left and right eye images into a single view is dependent on precise ocular alignment, which relies on coordinated eye movements. During movements of the head this alignment is maintained by numerous reflexes. Although rodents share with other mammals the key components of eye movement control, the coordination of eye movements in freely moving rodents is unknown. Here we show that movements of the two eyes in freely moving rats differ fundamentally from the precisely controlled eye movements used by other mammals to maintain continuous binocular fusion. The observed eye movements serve to keep the visual fields of the two eyes continuously overlapping above the animal during free movement, but not continuously aligned. Overhead visual stimuli presented to rats freely exploring an open arena evoke an immediate shelter-seeking behaviour, but are ineffective when presented beside the arena. We suggest that continuously overlapping visual fields overhead would be of evolutionary benefit for predator detection by minimizing blind spots.

Rats are commonly used as a model for studies of the mammalian visual system[1–4]. They have laterally facing eyes and a panoramic field of view extending in front, above and behind the animal's head[1]. Eye movements in head-restrained rats are conjugate[5], but studies of the vestibulo-ocular reflex in rats suggest that this only describes a fraction of their eye movements[6,7]. Rats can visually estimate distance for gap jumping[2,8] and perform object discrimination tasks[4], but in their natural environment also have to avoid predation from both airborne[9] and ground-dwelling predators[10]. This leads to conflicting demands on their visual system: on the one hand, maximum coverage of the environment for predator detection; on the other, detailed vision for object recognition and depth perception. Eye movements in freely moving rats have not been characterized so far, and in view of the conflicting pressures on their visual system it is unknown to what extent the trade-off between detailed vision and panoramic surveillance compromises their capacity for binocular fusion.

Eye movements in freely moving animals

To record eye movements in freely moving rats, we developed a miniaturized ocular-videography system that consisted of two lightweight head-mounted cameras (Supplementary Fig. 1). Pupil positions in the acquired images were tracked using custom-written algorithms. To allow analyses of the observed eye movements in the context of the rat's pose and location, we also tracked the position and orientation (pitch, roll and yaw) of the animal's head using a custom-built tracking system (see Supplementary Methods).

In freely moving animals, both eyes were highly mobile (Fig. 1a, b and Supplementary Video 1), with large horizontal and vertical excursions of the pupil (Fig. 1b). Both eyes moved continuously while the animal was exploring, but movements markedly reduced in amplitude when the animal stopped making large movements of its head. The dynamics of the movements were complex, regularly disconjugate and often asymmetrical. In addition to measuring horizontal and vertical pupil positions, we developed a method for tracking the irregular rough edge of the pupil in each frame which allowed measurement of ocular torsion (rotation around the optical axis) and quantification of torsional rotations (Fig. 1c, see Supplementary Methods). Torsional rotations occurred frequently, and reached relatively large amplitudes (20–30°; Fig. 1d and Supplementary Video 2). The dynamics of torsional rotations were also complex, and both cycloversion (rotation of both eyes in the same direction) and cyclovergence (rotation of the eyes in opposite directions) were observed (see *a* and *b* in Fig. 1d). On average there was a weak correlation between left and right eye torsion angles; however, the range of angles recorded for one eye for any given angle recorded for the other eye was very broad (Supplementary Fig. 2). In contrast to free movement, eyes movements in head-restrained rats were conjugate and infrequent, even when the animal was running on a spherical treadmill (Supplementary Fig. 3 and Supplementary Video 3).

Influence of head movements

Numerous sensory inputs and reflexes contribute to the regulation of eye position or gaze direction[6,11,12]. Particularly obvious in the current study was the role of the vestibulo-ocular reflex[6]. As previously observed in restrained rats, roll of the head to the right resulted in elevation of the right pupil and declination of the left pupil and vice versa for roll to the left (Fig. 2a, b). For both freely moving and head-restrained animals, these eye positions were maintained for as long as the roll was maintained (Supplementary Fig. 4 and Supplementary Video 4). Pitching of the head nose-up or -down resulted in strong convergent and divergent eye movements, respectively (Fig. 2c, d), and these positions were maintained while the pitch angle was maintained (Supplementary Fig. 4 and Supplementary Video 4). In addition, pitching of the head also resulted in complementary torsional rotation of the left and right eyes (Fig. 2e, f). To assess the extent to which the vestibulo-ocular reflex controlled the observed eye positions, we built a simple predictive model (see Supplementary Methods) which predicted eye positions based on pitch and roll of the head. The model was able to predict a large proportion of the tracked eye movements for both vertical ($78 \pm 2\%$ variance reduction, $n = 3$ animals) and horizontal axes ($69 \pm 3\%$ variance reduction, $n = 3$ animals; Supplementary Fig. 5).

[1]Network Imaging Group, Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany. [2]Bernstein Center for Computational Neuroscience Tübingen, Spemannstraße 41, 72076 Tübingen, Germany.

*These authors contributed equally to this work.

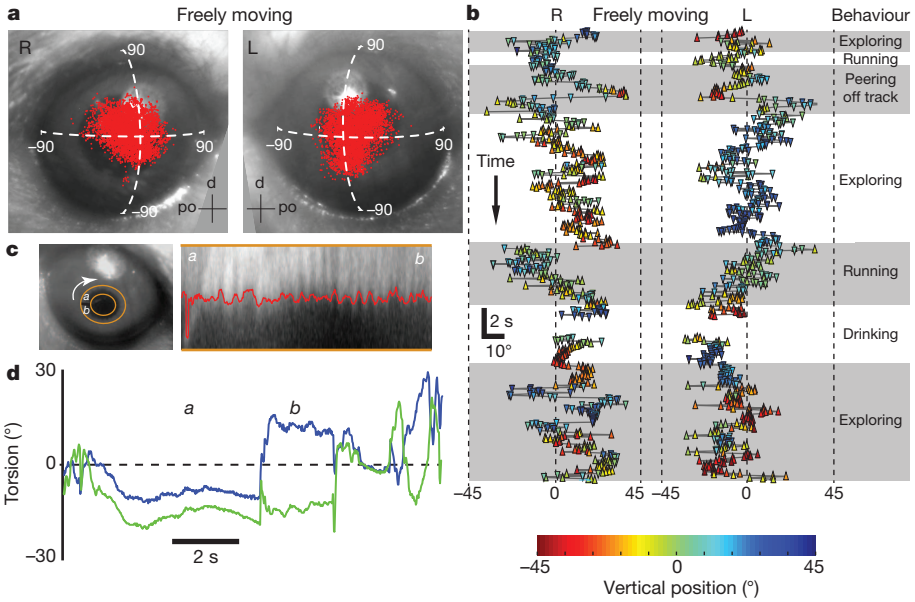


Figure 1 | Eye movements in freely exploring rats. **a**, Left and right eye images during free movement with individual pupil positions (red dots, approximately 5,000 data points) (dorsal (d) and posterior (po)). **b**, Vertical (marker colour) and horizontal (*x* axis position) kinetics (*y* axis) of eye movements during free movement (excerpt from **a**). Positive and negative vertical movements are denoted (up and down markers). Magnitude is represented (marker colour). Behavioural periods are indicated. **c**, Eye image (upper) showing the pupil margin used for torsional tracking (outlined in orange) and the extracted section (lower image) from upper image including tracked pupil margin (red). **d**, Torsion of right (green) and left (blue) eyes during free movement. Note eyes can both rotate in the same direction (*a*), opposite directions (*b*) and combinations thereof.

From this, we conclude that a large proportion of the eye movements we observed in freely moving animals were driven by vestibulo-ocular reflex.

Consequences for matching retinal images

One obvious feature of the observed eye movements was that the pointing directions of the two eyes often differed substantially (Supplementary Video 1). This observation implies that both the fraction of left and right eye retinal images that are matching and the location on the retina of any matching regions may vary from moment to moment. To begin to quantify this, we first measured the difference in pupil positions (right pupil position minus left pupil position; Fig. 3a and see Supplementary Methods for details). If this measure was used for animals with conjugate eye movements (human, primate, cat, etc.), differences in pupil positions would be minimal, other than during convergence and divergence. In the freely moving rat, the horizontal pupil position differences were both negative (one or both eyes rotating temporally away from the nose) and positive (convergent eye positions). This was also the case for the vertical plane, where positive differences represented a vertical divergence with the right eye more dorsal than the left, and vice versa for negative differences. The range of pupil position differences was large in both planes, with an average standard deviation of almost 20° (Supplementary Fig. 6). Furthermore, the differences in pupil positions in both planes changed continuously as the animal was moving (Supplementary Video 5), with the horizontal difference being strongly related to head pitch (Supplementary Fig. 7). In contrast, in head-restrained animals the differences in pupil positions were minimal (Fig. 3a), with the standard deviation nearly one-quarter that for freely moving animals (Supplementary Fig. 6). We also confirmed that these differences in pointing direction (gaze vectors) occurred when measured in a 'world coordinate' system (Fig. 3b; see Supplementary Methods and Supplementary Fig. 8) and the difference changed continuously, with shifts of more than 20° occurring several times per second (Fig. 3c).

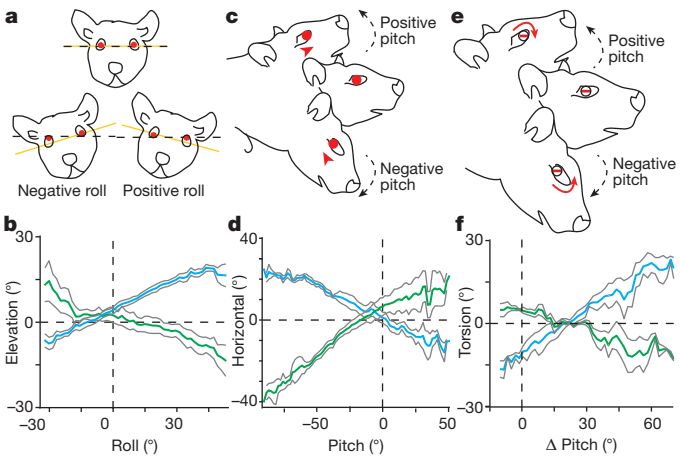


Figure 2 | Eye movements are dictated by head movement and position in freely moving animals. **a**, Schematic detailing how pupil elevation and depression (red pupils) can counteract head roll (yellow) compared with a horizon (black dashed). **b**, Comparison of pupil elevation for left (blue) and right (green) eyes in relation to head roll in a freely moving animal (average and s.e.m., $n = 4$ animals). **c**, Schematic detailing how eye movements in the horizontal plane (red arrowhead) occur during head pitch. **d**, Horizontal pupil position for left (blue) and right (green) eyes in relation to head pitch in a freely moving animal (average and s.e.m., $n = 4$ animals). **e**, Schematic detailing how ocular torsion (red arrows depict torsion direction) counteracts head pitch (black arrow) compared with horizon (red line). **f**, Ocular torsion for both left (blue) and right (green) eyes in relation to head pitch during free movement (average and s.e.m., $n = 4$ animals).

We next estimated the extent to which the observed eye movements may represent shifts in fixation onto different objects around the track as the animal performed a single cross of the gap. Because rats have no fovea or pronounced retinal specializations[13], measuring the extent to which fixation was maintained required an alternative reference point for re-projection over time. We therefore identified a time point shortly before the gap crossing when the animal's head position was at median pitch and roll, and then defined a reference visual target on the jumping track in the animal's field of view (Fig. 4a). Projection lines from this reference target into the centres of the left and right eyeballs were used to define the point on the surface of the eyeball to

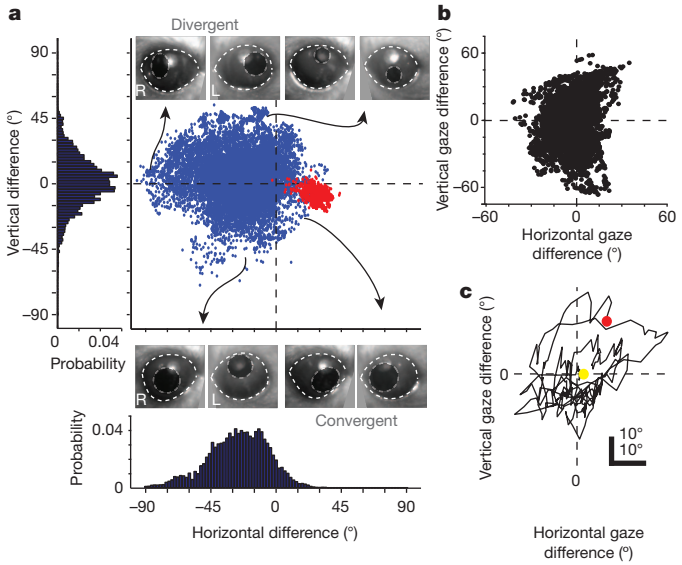


Figure 3 | Asymmetrical eye movements in freely moving rats. **a**, Distributions of the difference between left and right eye positions for a freely moving (blue) and head-restrained (red) rat. Each point represents the right eye position minus the left eye position for a single frame. Histograms are shown for *x* and *y* axes. Example image pairs (inset) from positions in the distribution (arrows). Conventions for eye images as in Fig. 1a. **b**, Scatter plot of the difference in left and right eye gaze vectors during free movement. **c**, Plot of the difference in left and right eye gaze vectors during free movement for a single continuous 1.7 s data segment including a gap cross.

be used for re-projection as the eye moved. To gauge the extent to which the observed ocular misalignment caused differences in potential visual targets of the two eyes, we rendered the environment around the rat, and followed the location where the re-projection lines contacted objects in the rendered environment (Fig. 4b, see Supplementary Methods). Over the 1.7 s required for the animal to perform the gap cross, most eye movements were disconjugate, resulting in a broad range of differences in both eye positions (Fig. 4c) and gaze vectors (Supplementary Fig. 9). The pupil projection points varied widely over the track (Fig. 4b), and there was very little coordination of the two points on single objects or locations (for rendered visualization see Supplementary Video 6). Note that the projections points

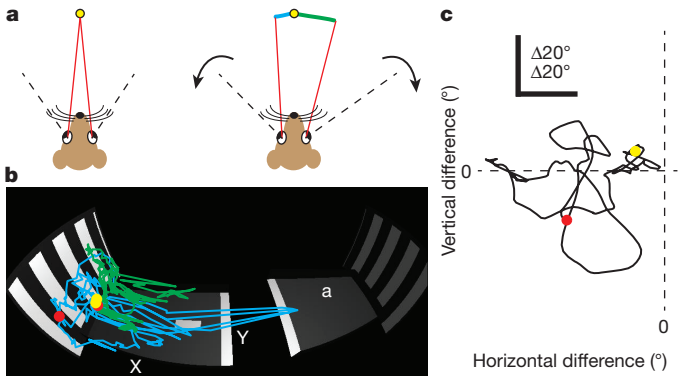


Figure 4 | Eye movements in freely moving animals are not consistent with those needed for binocular fusion. **a**, Schematic for defining lines of sight for re-projection. Left, reference visual target (yellow spot), optical axis (black), projections from visual target to eyeball centres (red). Right, relative changes of right (green) and left (blue) eye re-projections (red). **b**, Rendering of jumping arena showing monitors (far left and right stripes), initial animal position (a), initial gaze position (yellow dot for each eye) and subsequent gaze positions of the two eyes (left, green lines; right, blue lines; end gaze positions over 1.7 s ending with red dot). Same data as Fig. 3c. **c**, Difference between left and right eye positions for the data shown in **b** (conventions as Fig. 3a).

were precisely aligned on the reference visual target just before the jump. We next calculated the physical distance between the left and right eye projection points down the length and across the width of the track (Supplementary Fig. 9). In the animal's viewable environment, the distances separating the two projection points ranged from 0 to approximately 70 cm on the jumping track. Although we were not able to predict exactly what part of the visual space the animal was attending to, the constant changes in ocular alignment in both eye axes were not consistent with the animal shifting its gaze onto different objects of interest. We conclude that the coordination of eye movements in rats is not specialized for maintaining a fixed relationship between the eyes.

Maintenance of binocular field

The large collection angle of the rat eye (approximately 200°) combined with the lateral position of the eye on the head result in rats having large monocular visual fields, that share a large overlapping area extending in front, above and behind the animal's head[1] (Fig. 5a). To investigate the extent to which eye movements change the size, shape and location of the overlap of the monocular visual fields, we first generated a model of the animal's monocular visual fields based on optical and physiological properties of the rat eye[1]. The width of the overlapping fields at three different locations around the animal's head (Fig. 5b) varied strongly with the pitch of the animal's head (Fig. 5c, d and Supplementary Fig. 10). The width of the binocular field directly in front of the animal's nose, which is generally considered the animal's binocular viewing area[14], ranged from approximately 40° to 110° depending on head pitch. Changes in the extent of the visual field overlap measured at the inferior and posterior locations had strong but complementary dependence on head pitch

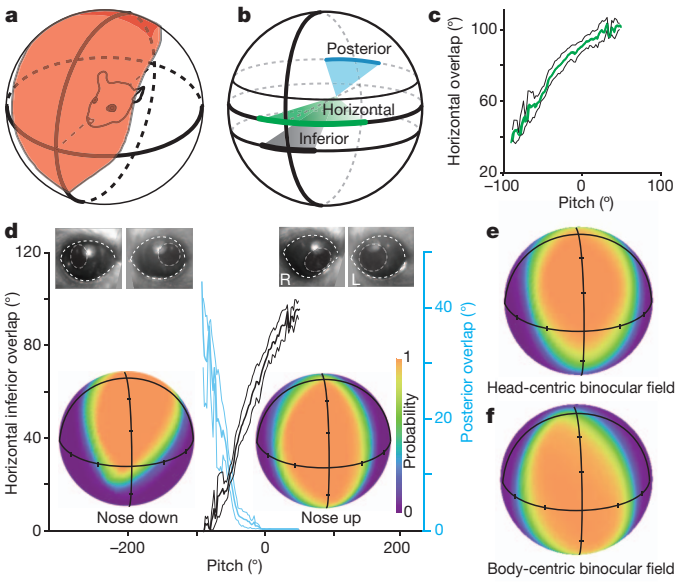


Figure 5 | Overhead binocular overlap. **a**, Schematic outlining binocular overlap (red, modified from ref. 1). **b**, Schematic for data in **c** and **d**. **c**, Average (green) dependence of horizontal overlap on head pitch (s.e.m., thin black lines, $n = 4$ animals). **d**, Dependence of horizontal inferior (black) and posterior (blue) overlap on head pitch (s.e.m., thin black lines, $n = 4$ animals). Head-centric density plots (insets) showing probability of visual field overlap (pseudo-colour) when animal is pitched down (≤10th centile of head pitch angles, insert left) or pitched up (≥90th centile, insert right, 30° ticks on vertical and horizontal axes). Note that average head roll was 18 ± 1° during nose-down pitch. Images (upper insets) show example eye positions for negative and positive head pitch (same as in Fig. 3a). **e**, Head-centric density plot of average overlap of monocular visual fields during free movement for all head positions (conventions as in **d**, $n = 4$ animals). **f**, Body-centric density plot of the overlapping fields that includes head and eye movements (conventions as in **d**, **e**, $n = 4$ animals). See Supplementary Fig. 11 for body-centric definition.

(Fig. 5d), consistent with the location of the binocular field remaining above the animal as the animal pitched its head. In all animals, the eye movements constantly kept the average overlap of the monocular visual fields above the animal's head (Fig. 5e). The effect of pitch on the location of this region was most clear when it was calculated for the top and bottom 10% of head pitch positions (average $-42.4 \pm 0.1^\circ$ for pitch down and $30.2 \pm 0.2^\circ$ pitch up; Fig. 5d, inserts). To characterize this further, we next calculated the position of the average binocular visual field relative to the animal's body (see Supplementary Fig. 11 for schematic). This 'bird's eye view' of the average overlap shows its location after accounting for the changing location of the visual fields caused by pitch and roll of the animal's head (Fig. 5f). In this reference system, the visual field overlap is predominantly located in front of and above the animal (Fig. 5f), despite an average nose-down head pitch of 25° (range 80° down to 40° up; Supplementary Fig. 11).

Together these results indicate that one of the key consequences of the eye movements observed in freely moving rats is that the region of overlap of the left and right visual fields is kept continuously above the animal, consistent with the suggestion that a major function of the rat visual system is to provide the animal with comprehensive overhead surveillance for predator detection[14].

Behavioural response to overhead stimuli

We next tested whether visual stimuli presented above the animal were capable of eliciting behavioural responses. Naive rats were placed in an open-field arena surrounded on three sides and above by stimulus monitors (Fig. 6a). The only object inside the open field was a shelter under which the animal could hide. Stimuli presented on the monitors beside the area failed to elicit any detectable changes in the animals' behaviour (Fig. 6b). In stark contrast, black moving stimuli presented overhead (Fig. 6c) elicited an immediate shelter-seeking behaviour from all animals tested (Fig. 6d and Supplementary Video 7). The rats ran immediately and directly to the shelter (Fig. 6e, 20 trials from three rats for side stimuli, 12 trials from three rats for overhead stimuli), and once there remained under the shelter for significantly extended periods (Fig. 6f, data sets as for Fig. 6e). As these behavioural responses may not necessarily require binocular viewing of the stimulus, one possibility is that the seemingly disconjugate eye movements, by continuously maintaining overlap of the monocular visual fields, help provide comprehensive surveillance of the region overhead by minimizing or eliminating 'blind spots'. However, it has also been shown for freely moving rats that certain aspects of their visual function, such as visual acuity, are enhanced in the binocular field compared with the monocular field[12]; thus it is also possible that these eye movements provide a direct enhancement of their vision by maintaining binocularity overhead. In summary, we conclude that although the observed eye movements preclude the possibility that rats continuously maintain binocular fusion while moving, they provide a benefit to the animal by facilitating comprehensive overhead surveillance as a defence against predation.

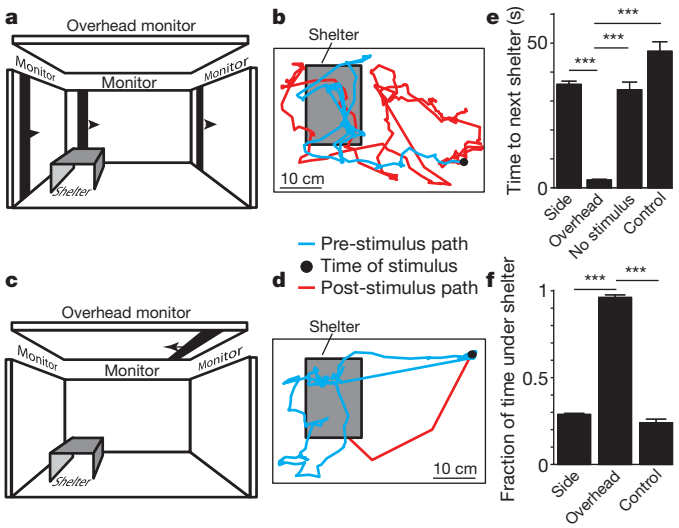


Figure 6 | Shapes moving overhead selectively evoke shelter-seeking behaviour. **a**, Schematic of side stimulus presentation. **b**, Animal's trajectory before (blue) and after (red) the onset (black circle) of a black moving bar stimulus presented on one of the side monitors. **c**, Schematic showing stimulus presentation above the rat. **d**, Trajectory before and after the onset of an overhead stimulus. Plot conventions as in **b**. **e**, Average time (bars, s.e.m.) before the rat's next visit underneath the shelter after stimulus presentation on monitors located beside the arena (Side), above the animal (Overhead), without stimulus presentation (No stimulus) or after a randomly chosen time in the data set (Control). **f**, Fraction of time spent underneath the shelter after stimuli presented on monitors beside the arena or overhead and for the same control condition described for **e**. Statistically significant group differences ($P < 0.01$) in **e** and **f** are denoted (stars, $n = 3$ animals).

Discussion

In primates, eye movements are precisely coordinated to maintain fixation of visual targets[15]. Precise ocular alignment is critical for binocular fusion. For foveal vision in humans misalignment of more than 1/3–1° results in double vision[16]. For peripheral vision, fusion is more tolerant to ocular misalignment; however, even there misalignment of more than a few degrees results in diplopia[17], and pupils moving in opposite vertical directions is associated with serious pathology[18]. In freely moving rats the difference in the gaze directions of the left and right eyes, which is a measure of the alignment of the eyes on a single target, has a range of more than 40° horizontally and more than 60° vertically. This range excludes the possibility that primate-like binocular fusion is continuously maintained when the animal is moving. Instead, eye movements in the rat are specialized for continuously maintaining overlap of the monocular visual fields above the animal as the head moves. It is clear from the low acuity[19], lack of fovea[13] and lack of significant capacity for accommodation[20] that rat vision is specialized along different lines to that of fovate mammals, and rats' strategy for eye movement control seems to be different as well. For the ground-dwelling rodent, foraging is actively pursued at dusk, and local changes in the environment are detected using mystacial vibrissa[21] and olfaction[22], both of which are associated with rapid head movements in all planes[23]. For rats, birds of prey such as owls[9] are a major predator, and as vision is the only sense that allows predator detection at a distance, the wide panoramic field of view[1,20], large depth of field[24] and maintenance of comprehensive overhead surveillance based on a system that counteracts the rapid head movements may be of substantial evolutionary advantage.

The eye movements observed here do not imply that rats are completely incapable of binocular fusion, stereoscopic depth perception or detailed vision. Rats can use their vision for depth perception[2,8] and are capable of quite sophisticated visual object recognition[4]. The variable alignment of the gaze directions of the eyes during head movements do imply, however, that for rats to fuse the two monocular images or to have stereoscopic depth perception they must either use a behavioural strategy to align the two monocular images (orient their head in a position that allows or facilitates fusion) or have another mechanism that allows them to identify matching components in the two retinal images. Some non-predatory bird species combine both panoramic vision (predator detection) with stereoscopic vision of close-by objects (bill vision) by using multiple retinal specializations[25], and other birds have behavioural strategies involving a combination of head movements for switching between distinct modes of viewing[26]. Rats may use similar strategies, in which the animal assumes a particular posture bringing both eye images into registration when detailed vision is required. An alternative proposal is that they can fuse left and right images without precise retinal

registration by using something like a corollary signal (for review, see ref. 27) to track the eye movements and identify matching retinal locations. This would be somewhat analogous to the mechanism suggested to explain shifting receptive field locations in monkey frontal cortex[27]. However, such a mechanism would require an immense degree of connectivity in the visual areas, and so far there is no evidence for this.

In summary, eye movements in freely moving rats are asymmetrical and inconsistent with the animal maintaining continuous fixation of a visual target with both eyes while moving. Instead, the movements keep the animal's binocular visual field above it continuously while it is moving, consistent with a primary focus of the animal's visual system being efficient detection of predators coming from above.

METHODS SUMMARY

The miniaturized camera system was secured onto a custom-built headplate which was implanted on the head. The position of the pupil was tracked in each image frame, and the effects of movement of the cameras eliminated by simultaneously tracking anatomical features of the eye (Supplementary Fig. 12 and Supplementary Video 8). The accuracy of the pupil-detection algorithm was measured to be less than 1°, and errors associated with tracking the anatomical features estimated to be very much less than 3° (Supplementary Fig. 13). Head position and orientation were tracked by following the relative position of six infrared light-emitting diodes mounted with the camera system. Tracking accuracy was less than 1° for all three axes of head orientation (Supplementary Fig. 14). For full details of all error quantifications, methods and analyses, see Supplementary Methods.

Received 10 December 2012; accepted 4 April 2013.
Published online 26 May 2013.

1. Hughes, A. A schematic eye for the rat. *Vision Res.* **19,** 569–588 (1979).
2. Legg, C. R. & Lambert, S. Distance estimation in the hooded rat: experimental evidence for the role of motion cues. *Behav. Brain Res.* **41,** 11–20 (1990).
3. Berardi, N. & Maffei, L. From visual experience to visual function: roles of neurotrophins. *J. Neurobiol.* **41,** 119–126 (1999).
4. Zoccolan, D., Oertelt, N., DiCarlo, J. J. & Cox, D. D. A rodent model for the study of invariant visual object recognition. *Proc. Natl Acad. Sci. USA* **106,** 8748–8753 (2009).
5. Chelazzi, L., Rossi, F., Tempia, F., Ghirardi, M. & Strata, P. Saccadic eye movements and gaze holding in the head-restrained pigmented rat. *Eur. J. Neurosci.* **1,** 639–646 (1989).
6. Hess, B. J. & Dieringer, N. Spatial organization of the maculo-ocular reflex of the rat: responses during off-vertical axis rotation. *Eur. J. Neurosci.* **2,** 909–919 (1990).
7. Quinn, K. J., Rude, S. A., Brettler, S. C. & Baker, J. F. Chronic recording of the vestibulo-ocular reflex in the restrained rat using a permanently implanted scleral search coil. *J. Neurosci. Methods* **80,** 201–208 (1998).
8. Russell, J. T. Depth discrimination in the rat. *Pedagog. Semin. J. Gen. Psychol.* **40,** 136–161 (1932).
9. Morris, P. Rats in the diet of the barn owl (*Tyto alba*). *J. Zool.* **189,** 540–545 (1979).
10. Doncanster, C. P., Dickman, C. R. & Macdonald, D. W. Feeding ecology of red foxes (*Vulpes vulpes*) in the city of Oxford, England. *J. Mamm.* **71,** 188–194 (1990).
11. Fuller, J. H. Eye and head movements in the pigmented rat. *Vision Res.* **25,** 1121–1128 (1985).
12. Prusky, G. T., Silver, B. D., Tschetter, W. W., Alam, N. M. & Douglas, R. M. Experience-dependent plasticity from eye opening enables lasting, visual cortex-dependent enhancement of motion vision. *J. Neurosci.* **28,** 9817–9827 (2008).
13. Euler, T. & Wassle, H. Immunocytochemical identification of cone bipolar cells in the rat retina. *J. Comp. Neurol.* **361,** 461–478 (1995).
14. Hughes, A. in *Handbook of Sensory Physiology* Vol. VII (ed. Crescitelli, F.) 613–756 (Springer, 1977).
15. Leigh, R. J. & Zee, D. S. *The Neurology of Eye Movement* 3rd edn (Oxford Univ. Press, 1999).
16. Duwaer, A. L. & van den Brink, G. What is the diplopia threshold? *Percept. Psychophys.* **29,** 295–309 (1981).
17. Lyle, T. K. & Foley, J. Subnormal binocular vision with special reference to peripheral fusion. *Br. J. Ophthalmol.* **39,** 474–487 (1955).
18. Dell'Osso, L. F. & Daroff, R. B. Two additional scenarios for see-saw nystagmus: achiasma and hemichiasma. *J. Neuro-Ophthalmol.* **18,** 112–113 (1998).
19. Douglas, R. M. *et al.* Independent visual threshold measurements in the two eyes of freely moving rats and mice using a virtual-reality optokinetic system. *Vis. Neurosci.* **22,** 677–684 (2005).
20. Hughes, A. The refractive state of the rat eye. *Vision Res.* **17,** 927–939 (1977).
21. Anjum, F., Turni, H., Mulder, P. G., van der Burg, J. & Brecht, M. Tactile guidance of prey capture in Etruscan shrews. *Proc. Natl Acad. Sci. USA* **103,** 16544–16549 (2006).
22. Wallace, D. G., Gorny, B. & Whishaw, I. Q. Rats can track odors, other rats, and themselves: implications for the study of spatial behavior. *Behav. Brain Res.* **131,** 185–192 (2002).
23. Munz, M., Brecht, M. & Wolfe, J. Active touch during shrew prey capture. *Front. Behav. Neurosci.* **4,** 191 (2010).
24. Artal, P., Herreros de Tejada, P., Munoz Tedo, C. & Green, D. G. Retinal image quality in the rodent eye. *Vis. Neurosci.* **15,** 597–605 (1998).
25. Fernandez-Juricic, E. *et al.* Testing the terrain hypothesis: Canada geese see their world laterally and obliquely. *Brain Behav. Evol.* **77,** 147–158 (2011).
26. Dawkins, M. S. What are birds looking at? Head movements and eye use in chickens. *Anim. Behav.* **63,** 991–998 (2002).
27. Wurtz, R. H. Neuronal mechanisms of visual stability. *Vision Res.* **48,** 2070–2089 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank W. Denk, R. Hahnloser, K. Kirchfeld, K. Martin, A. Schwartz and F. Wolf for comments on earlier versions of this manuscript and M. Rictis for help with electronics fabrication. We also thank A. Benali, A. Brauer, U. Czubayko, M.-L. Silva, V. Pawlak, V. Ramachandra and T. Senkova from the Network Imaging Group for data processing, A. Schaefer and M. Köllö for loan of the spherical treadmill, and F. Wolf for advice on the modelling and discussions throughout the project. We thank N. Logothetis for support and C. Sakmann for insights. G.N.'s salary was financed by the German Federal Ministry of Education and Research (BMBF; FKZ: 01GQ1002). The Max Planck Society financed research. We apologize to all authors whom we have not been able to cite because of space restrictions.

Author Contributions Experimental design was by D.J.W., D.S.G., J.S. and J.N.D.K., hardware and software development by D.S.G. and J.S., data collection by D.J.W., D.S.G., J.S. and S.R., analysis design and implementation by D.J.W., D.S.G., G.N., S.R. and J.N.D.K., methods text by D.J.W., D.S.G., J.S. and G.N. and composition of the main text by D.J.W. and J.N.D.K.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.N.D.K. (jason@tuebingen.mpg.de).

ARTICLE

doi:10.1038/nature12201

KAT5 tyrosine phosphorylation couples chromatin sensing to ATM signalling

Abderrahmane Kaidi[1] & Stephen P. Jackson[1,2]

The detection of DNA lesions within chromatin represents a critical step in cellular responses to DNA damage. However, the regulatory mechanisms that couple chromatin sensing to DNA-damage signalling in mammalian cells are not well understood. Here we show that tyrosine phosphorylation of the protein acetyltransferase KAT5 (also known as TIP60) increases after DNA damage in a manner that promotes KAT5 binding to the histone mark H3K9me3. This triggers KAT5-mediated acetylation of the ATM kinase, promoting DNA-damage-checkpoint activation and cell survival. We also establish that chromatin alterations can themselves enhance KAT5 tyrosine phosphorylation and ATM-dependent signalling, and identify the proto-oncogene c-Abl as a mediator of this modification. These findings define KAT5 tyrosine phosphorylation as a key event in the sensing of genomic and chromatin perturbations, and highlight a key role for c-Abl in such processes.

To maintain genome integrity is pivotal for cellular fitness[1], for which sensing genomic changes represents a critical step[2]. In response to DNA double-strand breaks (DSBs) within genomic DNA, chromatin organization is altered in an orchestrated manner to facilitate the cellular DNA damage response (DDR)[3,4]. One aspect of the DDR is checkpoint activation, which slows or halts cell-cycle progression[5]. Central to checkpoint signalling after DSBs is the serine-protein kinase ATM[6]. Although ATM is activated by its association with the MRE11–RAD50–NBS1 (MRN) complex at DSBs[7–10], accumulating evidence suggests that ATM activity is also potentiated by mechanisms that sense chromatin alterations in the context of DNA damage[11,12]. For example, binding of the protein lysine acetyltransferase KAT5 (also known as TIP60) to histone H3 trimethylated at lysine 9 (H3K9me3) promotes KAT5-dependent ATM acetylation, thereby enhancing ATM activity[12]. However, whether and how KAT5 binding to H3K9me3 is regulated has not been established. Here we show that the increase in KAT5 phosphorylation, mediated by the tyrosine kinase c-Abl, is involved in sensing chromatin alterations.

DNA-damage-dependent KAT5 Tyr phosphorylation

We first explored whether the KAT5–H3K9me3 interaction is regulated. Flag-tagged KAT5 was purified from cells before or after their exposure to ionizing radiation (Fig. 1a). As expected, we found that KAT5 bound an H3K9me3 peptide more effectively than a corresponding unmethylated peptide (Fig. 1b). Binding to H3K9me3 peptide was enhanced when KAT5 was purified from cells exposed to ionizing radiation (Fig. 1b; KAT5 failed to bind detectably to H3K4me3 and H3K27me3 peptides (Supplementary Fig. 1)), although this was reversed when KAT5 prepared from ionizing-radiation-treated cells was treated with λ-phosphatase (λ-PPase; Fig. 1b). These findings suggest that enhanced binding of KAT5 to H3K9me3 induced by ionizing radiation depends on KAT5 phosphorylation.

As KAT5 binds H3K9me3 through its chromodomain[12], we assessed this region for potential phosphorylation sites and identified a highly conserved residue, Tyr 44 (Fig. 1c). To test whether this site was phosphorylated, we expressed in human HeLa and RPE1 cells either Flag-tagged wild-type KAT5 (WT KAT5) or a derivative (YF KAT5) in which Tyr 44 was mutated to a non-phosphorylatable phenylalanine. After mock treatment or exposure to ionizing radiation, proteins were immunoprecipitated and probed with an anti-phospho-tyrosine antibody (pY). We observed ionizing-radiation-dependent tyrosine phosphorylation of KAT5; this modification required Tyr 44, as it was not detected with YF KAT5 protein (Fig. 1d). Phosphorylation of both recombinant and endogenous KAT5 occurred rapidly after ionizing radiation exposure (Fig. 1e, f). Although ionizing radiation increased the binding of WT KAT5 to H3K9me3, this was not the case for YF KAT5 protein (Fig. 1g).

Tyrosine phosphorylation enhances KAT5 activity

To assess the functional importance of KAT5 tyrosine phosphorylation, we examined whether mutating Tyr 44 affected the ability of an H3K9me3 peptide to stimulate KAT5-mediated ATM acetylation as detected by an acetyl-lysine (AcK) antibody[13]. An H3K9me3 peptide markedly stimulated ATM acetylation by WT KAT5 derived from ionizing-radiation-treated cells but not from non-irradiated cells (Fig. 2a and Supplementary Fig. 2a). By contrast, the peptide had little effect on YF KAT5 activity, irrespective of whether the protein was purified from cells exposed to ionizing radiation (Fig. 2a; no discernable acetyltransferase activity was observed with a catalytically inactive KAT5 derivative, KAT-I (Supplementary Fig. 2b)). Neither H3K9me3 peptide nor the KAT5 phosphorylation state affected the enzymatic activity of KAT5 on histone H4 (Fig. 2a), indicating that ionizing-radiation-dependent KAT5 tyrosine phosphorylation stimulates its acetylation of ATM specifically. Next, we established complementation systems in which HeLa or RPE cell lines contained a parental vector, or vectors, stably expressing WT KAT5 or YF KAT5 derivatives that are resistant to a short-interfering RNA (siRNA) that depletes the endogenous KAT5 protein. Consistent with our biochemical findings (Fig. 2a), the cellular defect in ionizing-radiation-induced ATM acetylation caused by KAT5 depletion was rescued by expression of WT KAT5 but not YF KAT5 (Fig. 2b). By contrast, both WT KAT5 and YF KAT5 proteins retained their housekeeping function, H4K16 acetylation within the *p73* promoter[14] (Fig. 2c). Consistent with the role of ATM acetylation in promoting its autophosphorylation on Ser 1981 (reported previously, see ref. 15), mutation of KAT5 Tyr 44 impaired ATM autophosphorylation in response to ionizing radiation (Fig. 2d).

[1]The Gurdon Institute and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. [2]The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK.

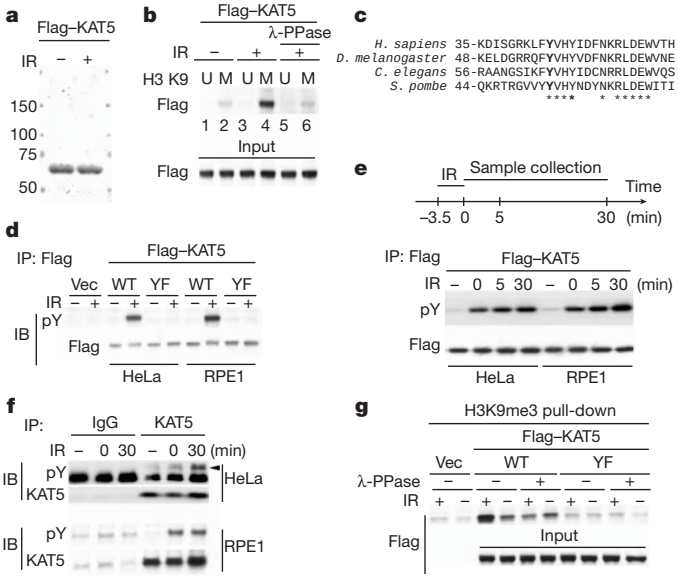


Figure 1 | KAT5 phosphorylation enhances its binding to H3K9me3. **a**, Flag–KAT5 was purified from HeLa cells and eluted with Flag peptides. **b**, Binding of Flag–KAT5 to H3-derived peptides (1–20) in which K9 is methylated (M) or unmethylated (U). **c**, Alignment of KAT5 chromodomains. Identical amino acid residues indicated with * symbol. **d**, Immunoprecipitations (IP) on extracts from cells expressing the indicated Flag–KAT5 constructs. **e**, Time course of Flag–KAT5 tyrosine phosphorylation after ionizing radiation (IR; 3 gray (Gy) in all cases). **f**, Time course of endogenous KAT5 tyrosine phosphorylation after IR (arrow indicates tyrosine-phosphorylated KAT5). **g**, Flag–KAT5 (WT KAT5 or YF KAT5) binding to H3K9me3 peptide. IB, immunoblot; IgG, immunoglobulin G.

These data therefore indicate that ionizing-radiation-induced KAT5 tyrosine phosphorylation is required for effective ATM activation, and suggest that YF KAT5 is a separation-of-function mutant that uncouples its DDR and housekeeping roles.

Using our complementation systems, we found that, unlike cells expressing endogenous or recombinant WT KAT5, those expressing YF KAT5 did not effectively phosphorylate ATM targets CHK2 and p53 after ionizing radiation (Fig. 2e). Cells expressing YF KAT5 also failed to trigger an ionizing radiation-induced G1/S DNA damage checkpoint (Fig. 2f and Supplementary Fig. 3; note the persistence of S phase cells in YF KAT5 compared to WT KAT5), and failed to arrest at the G2/M transition after ionizing radiation treatment (note the cells positive for the mitotic mark, histone H3 Ser-10 phosphorylation (H3S10p); Fig. 2g). While cells depleted of endogenous KAT5 and complemented with recombinant WT protein behaved like parental cells in clonogenic survival assays, those expressing YF KAT5 were hypersensitive to ionizing radiation (Fig. 2h). Collectively, these findings support a model in which ionizing radiation rapidly triggers the accumulation of tyrosine-phosphorylated KAT5, thereby promoting KAT5 binding to H3K9me3 and KAT5-mediated ATM acetylation, which in turn enhances ATM activation, ATM-mediated checkpoint signalling activation and cell survival.

Chromatin alterations induce KAT5 phosphorylation

Given that DNA damage can induce chromatin reorganization[16,17], our finding that KAT5 tyrosine phosphorylation connects H3K9me3-dependent chromatin binding to ATM activation suggests that this modification may constitute a sensing mechanism for chromatin changes. To test this idea, we altered chromatin in cells by inducing histone hyperacetylation with trichostatin A (TSA) or using siRNA-mediated depletion of heterochromatin protein 1α (HP1α; also known as chromobox protein homologue 5) to expose H3K9me3. Both TSA treatment and HP1α

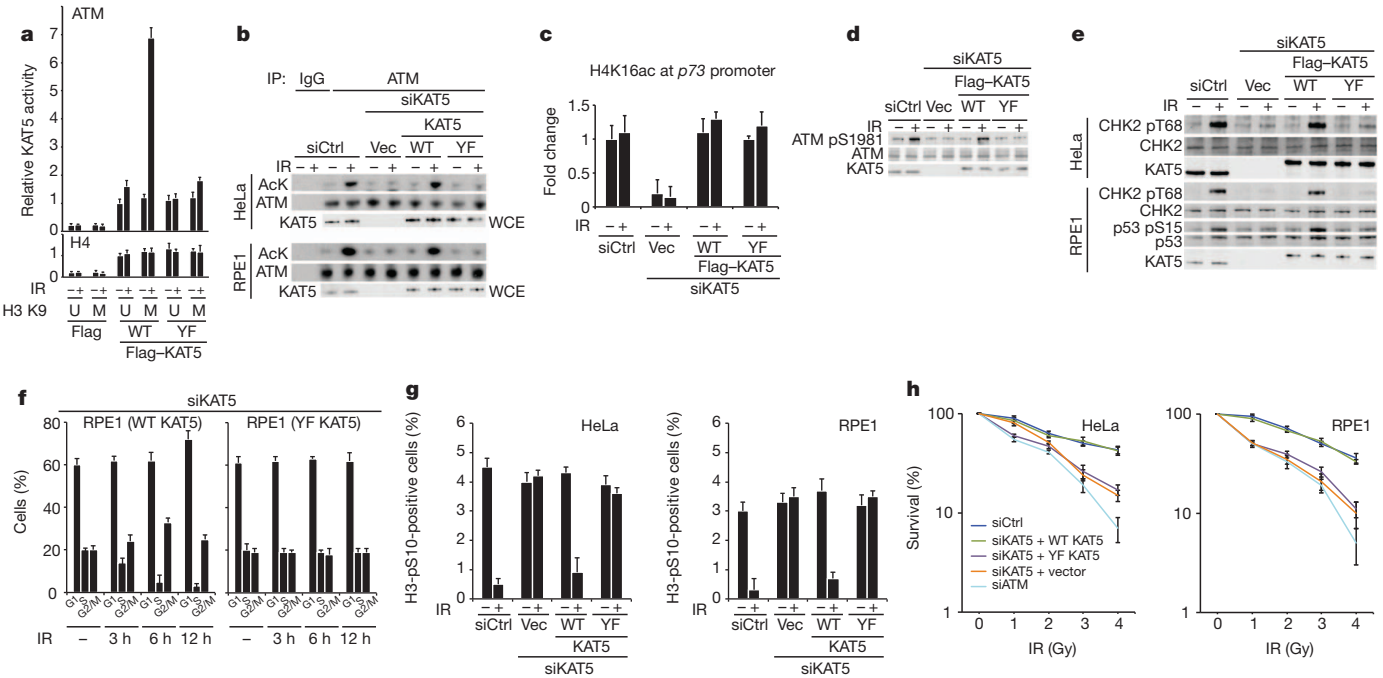


Figure 2 | KAT5 Tyr phosphorylation promotes ATM activation, checkpoint signalling and cell survival after ionizing radiation. **a**, Flag–KAT5 (WT KAT5 or YF KAT5) purified from HeLa cells was tested for ATM or H4 acetylation activity in the presence of H3K9me3 peptide. Data (mean ± s.e.) are from three experiments. **b**, ATM acetylation was examined after IP with an anti-acetyl-lysine (AcK) antibody. **c**, Analysis of *p73* promoter H4K16 acetylation by chromatin IP (ChIP). Data (means ± s.d.) were from three experiments. **d**, ATM Ser 1981 phosphorylation in RPE1 cell complementation system. **e**, ATM signalling in cells expressing siRNA-resistant KAT5 derivatives 1 h after 3 Gy IR. **f**, Cell-cycle analyses of RPE1 cells at indicated times after 3 Gy IR. Results (means ± s.e.m.) are from three experiments. **g**, G2/M DNA damage checkpoint assessed by measuring cells positive for H3 pS10 by flow cytometry 2 h post IR (3 Gy). Results (mean ± s.e.m.) are from three experiments. **h**, HeLa- and RPE1-cell survival after IR. Data (mean ± s.d.) are from three independent experiments. siCtrl, nonspecific siRNA control; siKAT5, siRNA against KAT5; Vec, empty vector; WCE, whole-cell extract.

depletion induced KAT5 tyrosine phosphorylation (Fig. 3a), ATM acetylation (Fig. 3b), ATM autophosphorylation (Supplementary Fig. 4) and ATM substrate phosphorylation (Fig. 3c and Supplementary Fig. 5a). Similarly, and consistent with a previous report[18], the DNA-intercalating agent chloroquine induced markers of ATM signalling (Supplementary Fig. 5b). Moreover, we found that these responses to TSA, HP1α depletion or chloroquine occurred with WT KAT5 but not YF KAT5 (Fig. 3c and Supplementary Fig. 5). ATM acetylation after TSA treatment was prevented by KAT5 Tyr 44 mutation, thus ruling out the possibility that TSA caused ATM acetylation by inhibiting a lysine deacetylase that targets ATM (Fig. 3d). Consistent with TSA triggering ATM activation in a manner that requires KAT5 tyrosine phosphorylation, we found that TSA caused cells to arrest in phase G1 and G2/M by mechanisms that were abrogated by KAT5 Tyr 44 mutation (Fig. 3e and Supplementary Fig. 6), or when cells were incubated with a selective ATM inhibitor (ATMi, KU-55933 (ref. 19); Fig. 3e and Supplementary Fig. 6). Cells expressing YF KAT5 that fail to induce cell cycle checkpoints in response to TSA treatment also displayed markedly reduced viability after acute TSA treatment (Fig. 3f).

The above data are consistent with chromatin alterations triggering KAT5 Tyr-44-dependent, ATM-mediated checkpoint activation. However, it was possible that TSA treatment or HP1α depletion resulted in DNA damage that activated KAT5 and ATM through canonical DDR mechanisms. This scenario seems unlikely because these treatments did not cause detectable DNA breaks (Supplementary Fig. 7). Furthermore, we found that siRNA-mediated depletion of MRE11, which is required for ATM activation by DSBs[7–10], abrogated ATM activation by ionizing radiation but had little effect on ATM activation in response to TSA treatment or HP1α depletion (Fig. 3g).

Chromatin binding fosters KAT5 phosphorylation

KAT5 Tyr 44 phosphorylation rapidly accumulates after treating non-irradiated cells with the phosphatase inhibitor okadaic acid (Fig. 4a and Supplementary Fig. 8; note that although okadaic acid induced ATM Ser 1981 phosphorylation at later times (3 h), it did not induce ATM acetylation). This suggests that KAT5 phosphorylation has a high turnover rate under normal conditions. We therefore speculated that this dynamic equilibrium of KAT5 phosphorylation may be altered after ionizing-radiation exposure if the phosphorylated form of KAT5 associated with perturbed chromatin, thus sequestering it from phosphatases and promoting its accumulation. Consistent with this idea,

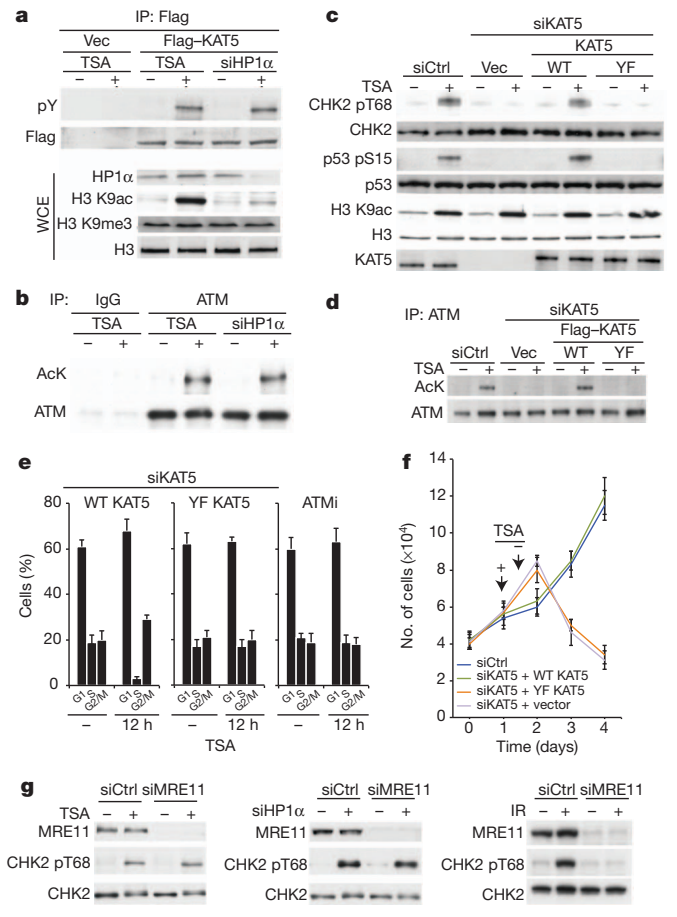


Figure 3 | Chromatin alterations activate KAT5 phosphorylation and checkpoint signalling. **a**, Flag–KAT tyrosine phosphorylation in RPE1 cells treated with TSA for 5 h or depleted for HP1α. **b**, ATM acetylation analysed after treatments as in **a**. **c**, **d**, RPE1 cells expressing siRNA-resistant KAT5 derivatives were examined for ATM-mediated signalling or ATM acetylation after 5 h TSA treatment. **e**, Flow-cytometry analyses of RPE1 cells treated with TSA for 12 h. Results (means ± s.e.m.) are from three experiments. **f**, RPE1 cells were treated with TSA for 16 h then cell proliferation was measured 24 h after TSA removal, as indicated. Data (mean ± s.e.m.) are from three experiments. **g**, MRE11 was depleted from RPE1 cells and cell extracts were analysed by western immunoblot analysis.

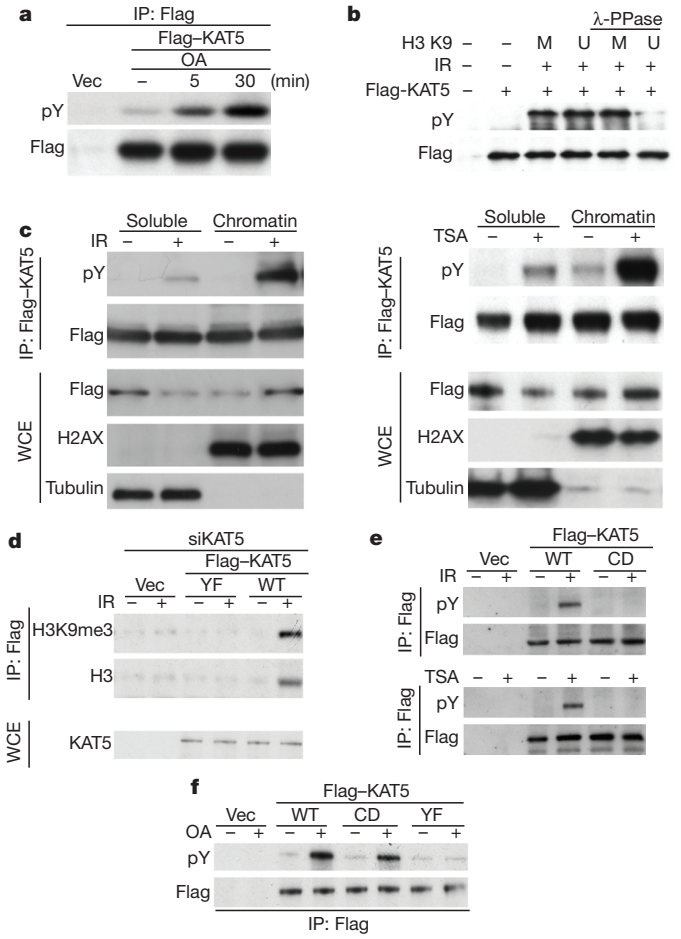


Figure 4 | Chromatin binding promotes KAT5 phosphorylation. **a**, RPE1 cells were treated with okadaic acid (OA; 25 nM). **b**, Flag–KAT5 from IR-treated cells was incubated with methylated (M) or unmethylated (U) H3-derived peptide and subjected to phosphatase treatment. **c**, KAT5 phosphorylation was examined after chromatin fractionation of RPE1 cells expressing Flag–KAT5 and treated with IR or TSA. H2AX, histone H2AX variant (unmodified). **d**, Flag-based IPs performed on extracts prepared from HeLa cells. **e**, WT and CD mutant Flag–KAT5 proteins were analysed for tyrosine phosphorylation after IP of extracts derived from IR- (upper panel) or TSA- (lower panel) treated RPE1 cells. **f**, WT, CD and YF Flag–KAT5 proteins were analysed for tyrosine phosphorylation after IP from mock- or OA-treated RPE1 cells.

addition of a methylated H3K9me3 peptide but not an unmethylated peptide protected KAT5 from dephosphorylation *in vitro* (Fig. 4b). Furthermore, cellular fractionation studies revealed that, although KAT5 was present in soluble and chromatin fractions, its tyrosine-phosphorylated form induced by ionizing radiation or TSA was predominantly chromatin-associated (Fig. 4c). In addition, co-immunoprecipitation studies revealed that, although WT KAT5 bound histone H3 and H3K9me3 in cells in an ionizing radiation-inducible manner, YF KAT5 failed to do so (Fig. 4d). Similarly, H3 bound by KAT5 was enriched for H3K36me3, another chromatin mark that simulates KAT5 activity *in vitro*[12] (Supplementary Fig. 9). Based on these observations and our finding that KAT5 binding to methylated H3 peptides impairs its dephosphorylation, we suspected that a KAT5 mutant defective in chromatin binding might fail to accumulate tyrosine phosphorylation. Indeed, a KAT5 variant bearing substitutions (F43A Y47A) in the chromodomain (CD) that abrogate H3K9me3 binding[13] (CD KAT5), displayed impaired ionizing radiation- and TSA-induced phosphorylation (Fig. 4e). Importantly, however, CD KAT5 Tyr 44 phosphorylation still accumulated after okadaic acid treatment, indicating that the CD mutation does not affect KAT5 phosphorylation but, instead, prevents its ability to be sequestered from phosphatase action upon chromatin alterations (Fig. 4f; note that okadaic acid did not induce phosphorylation of YF KAT5, highlighting Tyr 44 as a prime modification site). Together, these data support a model in which increased KAT5 tyrosine phosphorylation after exposure to ionizing radiation or chromatin perturbation arises through the induction of a permissive chromatin environment that is bound by phosphorylated KAT5, thereby enhancing its half-life and accumulation.

c-Abl targets KAT5 Tyr 44 to promote ATM signalling

We reasoned that the kinase(s) mediating KAT5 Tyr 44 phosphorylation would also probably have an effect on KAT5- and ATM-mediated DDR processes. Initial analyses ruled out a dependency of KAT5 phosphorylation on the DSB-responsive kinases ATM and DNA-dependent protein kinase (DNA-PK) (Supplementary Fig. 10). After assessing various other potential KAT5 kinases, we focused on the tyrosine kinase c-Abl, which has been implicated previously in DDR events[20–24]. Pre-treating cells with the small-molecule c-Abl inhibitor imatinib[25] (Gleevec) abolished c-Abl autophosphorylation on Tyr 245 (Fig. 5a, b; Supplementary Fig. 11 shows antibody validation; Supplementary Fig. 12 shows that this autophosphorylation was detectable in various cell lines and was not markedly affected by ionizing radiation or TSA treatment, or by ATM or DNA-PK inhibition, and that it is predominantly nuclear). Notably, imatinib also inhibited the accumulation of KAT5 phosphorylation as well as CHK2 phosphorylation after ionizing radiation or TSA treatment (Fig. 5a, b). Inhibition of c-Abl also prevented ionizing-radiation-induced KAT5 chromatin accumulation and H3K9me3 binding (Supplementary Fig. 13). Moreover, pre-treating cells with imatinib abrogated the G1/S cell-cycle checkpoint after exposure to ionizing radiation or TSA (Fig. 5c and Supplementary Fig. 14a) and interfered with the induction of the G2/M checkpoint in response to ionizing radiation or TSA exposure (Fig. 5d and Supplementary Fig. 14b).

The finding that siRNA-mediated depletion of c-Abl also abolished the accumulation of KAT5 tyrosine phosphorylation after ionizing radiation or TSA treatment suggests that the above effects are not caused by the inhibition of other kinases by imatinib (Fig. 5e, f). Moreover, c-Abl depletion markedly impaired the ability of ionizing radiation and TSA to trigger ATM-mediated phosphorylation of CHK2 and p53 (Fig. 5e, f and Supplementary Fig. 15a; consistent with previous findings[22]), and abrogated ionizing-radiation-induced checkpoint responses (Supplementary Fig. 15b). To assess whether these effects reflect direct targeting of KAT5 by c-Abl, we carried out *in vitro* kinase experiments. Thus, we found that purified c-Abl mediated tyrosine phosphorylation of purified WT KAT5 but not YF KAT5 (Fig. 5g; note that 22 tyrosine residues remain within YF KAT5, indicating that c-Abl is not a non-specific kinase under our assay conditions). Furthermore, pre-phosphorylation of KAT5 by c-Abl enhanced ATM acetylation by KAT5 in an H3K9me3-dependent manner (Fig. 5h).

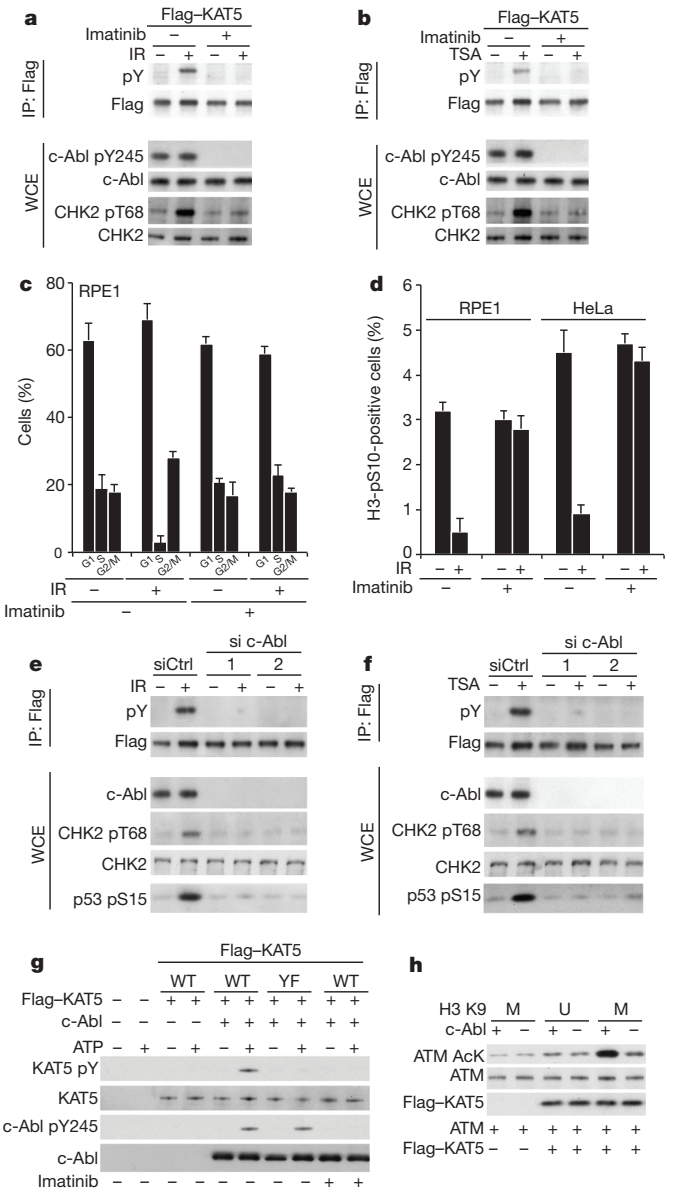


Figure 5 | c-Abl-dependent KAT5 tyrosine phosphorylation. **a**, **b**, RPE1 cells were treated with imatinib (1 μM) for 3 h and then exposed to IR (**a**) or TSA treatment (**b**). **c**, Flow-cytometry analyses of RPE1 cells exposed to IR in the presence or absence of imatinib. Cells were analysed after 8 h (data: means ± s.e.m., $n = 3$). **d**, Cells were treated, collected 2 h post IR and analysed for H3pS10 by flow cytometry (data: mean ± s.e.m.; $n = 3$). **e**, **f**, RPE1 cells were transfected with c-Abl targeting siRNAs, and exposed to IR (**e**) or TSA (**f**), then analysed. **g**, *In vitro* kinase assays performed with recombinant c-Abl and purified Flag-KAT5 (WT KAT5 or YF KAT5). Imatinib treatment at 25 nM concentration. **h**, Flag–KAT5 purified from HeLa cells was bound to beads and then subjected to c-Abl-mediated phosphorylation. After washing, its activity towards ATM was assessed in the presence of K9 methylated (M) or unmethylated (U) H3 peptides.

Discussion

We have identified a mechanism involving tyrosine phosphorylation of KAT5 that underlies cellular sensing of chromatin perturbations and its coupling to checkpoint signalling. As DNA lesions result in local and global chromatin remodelling[16,17], this study highlights how

sensing of chromatin alterations associated with DNA lesions has a major role in promoting an effective DDR. Although the association of KAT5 with H3K9me3 may suggest that KAT5 activity would be most pronounced in heterochromatin, H3K9me3 is found within other loci[26], and KAT5 can also be activated by H3K36me3 (ref. 12), which has a broad genomic distribution[27,28]. Our data indicate that chromatin alterations can instigate checkpoint signalling independently of DNA breaks; thus, proper chromatin organization is probably monitored continuously within the cell. We have found that perturbed chromatin promotes KAT5 chromatin binding and the concurrent accumulation of KAT5 Tyr 44 phosphorylation, which in turn induces ATM-mediated signalling and checkpoint activation. These results and our findings linking c-Abl to KAT5 help to explain earlier observations of ATM activation after chromatin alterations[11,18]. In addition, they extend previous studies linking c-Abl to DDR processes[21–24] and describing functional interactions between c-Abl and KAT5 (ref. 29).

Although our data imply that basal c-Abl kinase activity[30] is sufficient for the accumulation of tyrosine-phosphorylated KAT5 after the generation of a permissive chromatin environment, it will be of interest to see whether enhancements of c-Abl activity that have been reported to occur in response to DNA damage[20,24] serve to potentiate ATM signalling in certain settings. It is also tempting to speculate that c-Abl- and KAT5-dependent ATM and DDR signalling may be dysregulated in progeria cells, where heterochromatin organization is perturbed[31–33]. Finally, our findings suggest new avenues for therapeutic intervention. They may explain the mode of action of certain lysine deacetylase inhibitors that are being developed as therapeutic agents[34], and may provide insight into how such drugs can be best employed against cancer and other diseases. Our work also highlights how drugs targeting c-Abl may potentiate the effects of chromatin-modulating drugs, radiotherapy and DNA-damaging chemotherapies[35], and may also have potential in situations in which cancer cells are particularly reliant on KAT5- and ATM-mediated signalling because they possess inherent DDR defects or are subject to ongoing DNA damage and/or chromatin perturbations.

METHODS SUMMARY

Stable cells expressing tagged KAT5 were selected in G418 (1 mg ml^{-1}). For Flag-based protein purifications, cell lysates were prepared with benzonase and then high salt (450 mM) extraction. After immunoprecipitation, complexes were eluted with 3× Flag peptide. For chromatin fractionations, the soluble fraction was obtained with cytoskeleton (CSK) buffer, and subsequent to this the chromatin fraction was recovered with benzonase and then high salt extraction. Peptide binding assays were carried out as described previously[12] (see also Methods). Lysine acetyl-transferase assays were carried out as described previously[12] but with modifications as described in the Methods. Checkpoint assays involved flow cytometry on fixed cells, either by staining for DNA content alone or in combination with staining for histone H3S10p. Full methods and associated references are available in the online version of this paper.

Full Methods and any associated references are available in the online version of the paper.

Received 4 December 2012; accepted 18 April 2013.
Published online 26 May 2013.

1. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461,** 1071–1078 (2009).
2. Harper, J. W. & Elledge, S. J. The DNA damage response: ten years after. *Mol. Cell* **28,** 739–745 (2007).
3. Miller, K. M. & Jackson, S. P. Histone marks: repairing DNA breaks within the context of chromatin. *Biochem. Soc. Trans.* **40,** 370–376 (2012).
4. Downs, J. A., Nussenzweig, M. C. & Nussenzweig, A. Chromatin dynamics and the preservation of genetic information. *Nature* **447,** 951–958 (2007).
5. Bartek, J. & Lukas, J. DNA damage checkpoints: from initiation to recovery or adaptation. *Curr. Opin. Cell Biol.* **19,** 238–245 (2007).
6. Shiloh, Y. The ATM-mediated DNA-damage response: taking shape. *Trends Biochem. Sci.* **31,** 402–410 (2006).
7. Lee, J.-H. & Paull, T. T. Direct activation of the ATM protein kinase by the Mre11/Rad50/Nbs1 complex. *Science* **304,** 93–96 (2004).
8. Lee, J.-H. & Paull, T. T. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science* **308,** 551–554 (2005).
9. Falck, J., Coates, J. & Jackson, S. P. Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature* **434,** 605–611 (2005).
10. Uziel, T. *et al.* Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J.* **22,** 5612–5621 (2003).
11. Ayoub, N., Jeyasekharan, A. D., Bernal, J. A. & Venkitaraman, A. R. HP1-beta mobilization promotes chromatin changes that initiate the DNA damage response. *Nature* **453,** 682–686 (2008).
12. Sun, Y. *et al.* Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60. *Nature Cell Biol.* **11,** 1376–1382 (2009).
13. Sun, Y., Xu, Y., Roy, K. & Price, B. D. DNA damage-induced acetylation of lysine 3016 of ATM activates ATM kinase activity. *Mol. Cell. Biol.* **27,** 8502–8509 (2007).
14. Kimura, A. & Horikoshi, M. Tip60 acetylates six lysines of a specific class in core histones *in vitro*. *Genes Cells* **3,** 789–800 (1998).
15. Sun, Y., Jiang, X., Chen, S., Fernandes, N. & Price, B. D. A role for the Tip60 histone acetyltransferase in the acetylation and activation of ATM. *Proc. Natl Acad. Sci. USA* **102,** 13182–13187 (2005).
16. Ziv, Y. *et al.* Chromatin relaxation in response to DNA double-strand breaks is modulated by a novel ATM- and KAP-1 dependent pathway. *Nature Cell Biol.* **8,** 870–876 (2006).
17. Kruhlak, M. J. *et al.* Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *J. Cell Biol.* **172,** 823–834 (2006).
18. Bakkenist, C. J. & Kastan, M. B. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. *Nature* **421,** 499–506 (2003).
19. Hickson, I. *et al.* Identification and characterization of a novel and specific inhibitor of the ataxia-telangiectasia mutated kinase ATM. *Cancer Res.* **64,** 9152–9159 (2004).
20. Shafman, T. *et al.* Interaction between ATM protein and c-Abl in response to DNA damage. *Nature* **387,** 520–523 (1997).
21. Meltser, V., Ben-Yehoyada, M. & Shaul, Y. c-Abl tyrosine kinase in the DNA damage response: cell death and more. *Cell Death Differ.* **18,** 2–4 (2011).
22. Wang, X. *et al.* A positive role for c-Abl in Atm and Atr activation in DNA damage response. *Cell Death Differ.* **18,** 5–15 (2011).
23. Kharbanda, S. *et al.* Activation of the c-Abl tyrosine kinase in the stress response to DNA-damaging agents. *Nature* **376,** 785–788 (1995).
24. Baskaran, R. *et al.* Ataxia telangiectasia mutant protein activates c-Abl tyrosine kinase in response to ionizing radiation. *Nature* **387,** 516–519 (1997).
25. Buchdunger, E. *et al.* Inhibition of the Abl protein-tyrosine kinase *in vitro* and *in vivo* by a 2-phenylaminopyrimidine derivative. *Cancer Res.* **56,** 100–104 (1996).
26. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
27. Edmunds, J. W., Mahadevan, L. C. & Clayton, A. L. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *EMBO J.* **27,** 406–420 (2008).
28. Chantalat, S. *et al.* Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res.* **21,** 1426–1437 (2011).
29. Jiang, Z. *et al.* Tip60-mediated acetylation activates transcription independent apoptotic activity of Abl. *Mol. Cancer* **10,** 88 (2011).
30. Brasher, B. B. & Van Etten, R. A. c-Abl has high intrinsic tyrosine kinase activity that is stimulated by mutation of the Src homology 3 domain and by autophosphorylation at two distinct regulatory tyrosines. *J. Biol. Chem.* **275,** 35631–35637 (2000).
31. Scaffidi, P. & Misteli, T. Lamin A-dependent nuclear defects in human aging. *Science* **312,** 1059–1063 (2006).
32. Pegoraro, G. & Misteli, T. The central role of chromatin maintenance in aging. *Aging (Albany NY)* **1,** 1017–1022 (2009).
33. Pegoraro, G. *et al.* Ageing-related chromatin defects through loss of the NURD complex. *Nature Cell Biol.* **11,** 1261–1267 (2009).
34. Lane, A. A. & Chabner, B. A. Histone deacetylase inhibitors in cancer therapy. *J. Clin. Oncol.* **27,** 5459–5468 (2009).
35. Podtcheko, A. *et al.* Inhibition of ABL tyrosine kinase potentiates radiation-induced terminal growth arrest in anaplastic thyroid cancer cells. *Radiat. Res.* **165,** 35–42 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all members of the Jackson laboratory for help and support, and A. Blackford, S. Britton, K. Dry, J. Forment and J. Travers for critical reading of the manuscript. Research in the Jackson laboratory is funded by Cancer Research UK program grant C6/A11224, the European Research Council and the European Community Seventh Framework Programme grant agreement no. HEALTH-F2-2010-259893 (DDR). Core funding is provided by CRUK (C6946/A14492) and the Wellcome Trust (WT092096). S.P.J. receives his salary from the University of Cambridge, UK, supplemented by CRUK. A.K. is funded by a Herchel Smith Fellowship from the University of Cambridge.

Author Contributions All experiments were conceived by A.K. and S.P.J. and were carried out by A.K. S.P.J. and A.K. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.P.J. (s.jackson@gurdon.cam.ac.uk).

METHODS

Cell culture and transfections. HeLa cells were grown in Dulbecco's modified Eagle medium (DMEM, Sigma-Aldrich) supplemented with 10% fetal bovine serum (BioSera), 2 mM L-glutamine, 100 U per ml penicillin, 100 μg ml^{-1} streptomycin and fungizone (Sigma-Aldrich). RPE1 cells were grown in DMEM and Ham F12 mix medium supplemented as above and buffered with sodium bicarbonate. Stably transfected HeLa and RPE1 cells were maintained in standard medium containing 1 mg ml^{-1} G418 (Invitrogen). The siRNA duplexes were obtained from MWG-Biotech (Supplementary Table 2). Plasmid DNA and siRNA transfections were carried out using Lipofectamine 2000 and Lipofectamine RNAiMax (Invitrogen), respectively, following the manufacturer's instructions. Cells were analysed 48 to 72 h after transfection.

DNA damage and drug treatments. ATM inhibitor (ATMi, KU-55933) was provided by KuDOS Pharmaceuticals (AstraZeneca), and used at a concentration of 10 μM, with 1 h pre-treatment. Trichostatin A (TSA, Cell Signaling Technology) was used at 1.3 μM for either 5, 12 or 16 h, as indicated. Okadaic acid (AbCam) was used at 25 nM for 5 or 30 min. Imatinib (Enzo Life Sciences) was used at a concentration 1 μM with a 3-h pre-treatment. Ionizing radiation was delivered by an X-ray generator (Faxitron X-ray Corporation RX-650; 120 kV, 5 mA, dose rates of 10 and 0.86 Gy per min).

Flow cytometry. Cells were fixed in ice-cold 70% ethanol. DNA was stained with 50 μg ml^{-1} propidium iodide (Sigma-Aldrich) in phosphate buffer solution (PBS) containing 0.1% Triton-X-100 and 0.5 mg ml^{-1} DNase free RNase A (Sigma-Aldrich). Samples were processed on a FACSCalibur flow cytometer equipped with CellQuest software (Becton Dickinson). Results were analysed using FlowJo software (TreeStar).

TUNEL assays. TdT-mediated dUTP nick end labelling (TUNEL) assays were carried out using a DeadEnd Fluorometric TUNEL System Kit (Promega), using the manufacturer's protocol. Cells grown on glass coverslips (VWR) were fixed with 4% paraformaldehyde in PBS for 20 min at room temperature (20 to 25 °C). After three washes in PBS, cells were permeabilized with 0.2% Triton-X-100 in PBS for 5 min at room temperature. Coverslips were equilibrated in 100 μl of equilibration buffer for 10 min. After this, label was washed away using 50 μl of TdT reaction mix for 1 h at 37 °C in the dark. The labelling reaction was stopped using 2× SSC buffer for 15 min then washed three times in PBS. Coverslips were mounted on glass slides using mounting medium containing DAPI (4′,6-diamidino-2-phenylindole). Confocal images were viewed using the Olympus FluoView1000 system. To avoid bleed-through effects in double-staining experiments, each dye was scanned independently in a multi-tracking mode. Samples were scanned using an ×40 or ×60 oil objective.

Comet assay. Neutral comet assays were carried out as specified by the Comet Assay kit (Trevigen) using GelBond films (Lonza) to support agarose gels. Samples stained with SYBR-Green I were observed under an epifluorescence microscope (Olympus IX71) using a UPlanFLN 10× objective. Images were analysed with CometScore software (TriTek) by scoring approximately 100 cells in each case.

Immunoblot analysis. Total cell extracts were prepared by scraping cells in Laemmli buffer (0.8% SDS, 4% glycerol, 280 mM β-mercaptoethanol, 25 mM Tris-HCl, pH 6.8). Proteins were resolved by SDS–PAGE (SDS–polyacrylamide gel electrophoresis), transferred onto nitrocellulose membrane (Protran), and probed using the appropriate primary (Supplementary Table 2) and secondary antibodies coupled to horseradish peroxidase (HRP; Dako-Pierce). Detection was carried out with ECL Western Blotting detection reagent (GE Healthcare).

Immunoprecipitation and protein purification. Cells were collected in PBS and lysed for 5 min at room temperature in immunoprecipitation lysis buffer (20 mM Tris-HCl, pH 7.5, 40 mM NaCl, 2 mM $MgCl_2$, 0.5% NP-40) freshly supplemented with 50 U per ml benzonase (Roche) and EDTA-free protease and phosphatase inhibitor cocktails (Roche). After this initial lysis step, NaCl concentration was adjusted to 450 mM and samples were incubated at 4 °C with rotation. Lysates were clarified by centrifugation (16,000 *g*, 20 min at 4 °C), and after recovery NaCl concentration was equilibrated to 150 mM. Protein concentration was then determined and lysates of 1 or 2 mg protein were used for immunoprecipitation reactions in immunoprecipitation buffer (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1.5 mM DTT, 10% glycerol, 0.5% NP-40) supplemented with protease and phosphatase inhibitors. Target proteins were captured with the appropriate antibody and protein A/G-sepharose Fast-Flow (Sigma) or protein A/G-Dynabeads (Dynal). Complexes were washed five times with immunoprecipitation buffer supplemented with protease and phosphatase inhibitors. Immunoprecipitation with rabbit serum, mouse serum, haemagglutinin, or from cells that do not express epitope-tagged protein were used as negative controls. For Flag-based purification, a similar protocol was used, but purifications were carried out using anti-Flag-M2 beads (Sigma) followed by elution with 3× Flag peptide (Sigma) in TBS buffer containing protease and phosphatase inhibitors, following the manufacturer's protocol.

Chromatin fractionation. Cells were treated with cytoskeleton buffer (10 mM PIPES, pH 7.0, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 0.5% Triton-X-100 with protease and phosphatase inhibitors) for 10 min at 4 °C. This soluble fraction was collected and cleared by centrifugation (13,200 r.p.m., 20 min at 4 °C). The CSK-resistant (chromatin fraction) was prepared using a method similar to the sample preparation for immunoprecipitation described above.

Peptide binding assays. Peptide binding assays were carried out as described previously[12]. Accordingly, biotinylated peptides (5 μg) were pre-bound to sepharose-avidin beads for 3 h at 4 °C, and washed extensively in binding buffer (20 mM HEPES-KOH, pH 7.5, 150 mM NaCl, 2.5% glycerol, 0.05% Tween 20, 2 mM DTT, supplemented with protease and phosphatase inhibitor cocktail (Roche)). Flag–KAT5 eluates were incubated with peptide bound to sepharose-avidin beads for 2 h in binding buffer at 4 °C. Beads were washed with wash buffer (20 mM HEPES-KOH, pH 7.9, 300 mM KCl, 0.2% Tween 20, 1.5 mM $MgCl_2$, 2 mM DTT, supplemented with protease and phosphatase inhibitor cocktail (Roche)). Bound proteins were eluted from the resin with SDS sample buffer and analysed by SDS–PAGE.

KAT assays. For lysine acetyltransferase (KAT) assays, Flag–KAT5 eluates (prepared as described above) were initially dialysed in KAT assay buffer (50 mM Tris-HCl, pH 8.0, 10% glycerol, 0.1 mM EDTA and 1 mM DTT containing protease and phosphatase inhibitors), and protein concentrations were determined. KAT reactions were performed in 60 μl of KAT assay buffer containing 0.5 μg of Flag-KAT5 eluate and 1 μg of purified ATM (provided by KuDOS Pharmaceuticals) or 1 μg of recombinant histone H4 (AbCam) in the presence of acetyl coenzyme A (100 μM) for 30 min at 30 °C. Samples were then separated by SDS–PAGE and examined by quantitative western blot analysis using an acetyl-lysine antibody (AcK).

***In vitro* kinase assay.** *In vitro* kinase assays with c-Abl were carried out in 30 μl of a buffer containing 50 mM Tris-HCl, pH 7.5, 5 mM $MnCl_2$, 0.25 μg of recombinant c-Abl (Enzo Life Sciences) and 0.5 μg of KAT5 eluate. The kinase reactions were initiated by the addition of 50 μM ATP and performed at 30 °C. The reactions were terminated after 30 min by addition of 5× protein sample buffer and boiling for 5 min. The samples were then separated using SDS–PAGE and examined by western blot analysis.

Colony forming assay. Forty-eight hours after siRNA transfection, cells were replated and exposed to ionizing radiation the next day. Subsequently, cells were incubated for an additional 14 days at 37 °C to allow colony formation. Colonies were stained with 0.5% crystal violet solution in 20% ethanol, and then counting was performed. Results were normalized to plating efficiencies.

LETTER

doi:10.1038/nature12165

Bounding the pseudogap with a line of phase transitions in $YBa_2Cu_3O_{6+\delta}$

Arkady Shekhter[1], B. J. Ramshaw[1], Ruixing Liang[2,3], W. N. Hardy[2,3], D. A. Bonn[2,3], Fedor F. Balakirev[1], Ross D. McDonald[1], Jon B. Betts[1], Scott C. Riggs[4,5] & Albert Migliori[1]

Close to optimal doping, the copper oxide superconductors show 'strange metal' behaviour[1,2], suggestive of strong fluctuations associated with a quantum critical point[3–6]. Such a critical point requires a line of classical phase transitions terminating at zero temperature near optimal doping inside the superconducting 'dome'. The underdoped region of the temperature–doping phase diagram from which superconductivity emerges is referred to as the 'pseudogap'[7–13] because evidence exists for partial gapping of the conduction electrons, but so far there is no compelling thermodynamic evidence as to whether the pseudogap is a distinct phase or a continuous evolution of physical properties on cooling. Here we report that the pseudogap in $YBa_2Cu_3O_{6+\delta}$ is a distinct phase, bounded by a line of phase transitions. The doping dependence of this line is such that it terminates at zero temperature inside the superconducting dome. From this we conclude that quantum criticality drives the strange metallic behaviour and therefore superconductivity in the copper oxide superconductors.

Resonant ultrasound spectroscopy (RUS) measures the frequencies f_n and widths Γ_n of the vibrational normal modes of a crystal acting as a free mechanical resonator. The frequencies of the normal modes are determined by the density and geometry of the crystal as well as by its elastic properties. The elastic component of the temperature evolution of these frequencies, $\Delta f_n(T)$, depends on a linear combination of all elastic moduli and reflects changes in the thermodynamic state of the system such as those associated with a phase transition. The width of a resonance, $\Gamma_n(T)$, is proportional to the energy dissipation caused by time-dependent (dynamic) fluctuations in the system. Measuring many resonances provides access to elastic properties and fluctuations with different symmetries[14–17]. Recent advances in the quality of single crystal $YBa_2Cu_3O_{6+\delta}$ (YBCO) have pushed the boundary of possible measurements, as demonstrated by the observation of quantum oscillations[18]. Advances in resonant ultrasound enable the determination of the thermodynamics of these submillimetre crystals to an accuracy of parts per million.

The narrow temperature range over which the resonances evolve across the superconducting transition illustrates the quality of the crystals and the accuracy of the measurement[19] (Fig. 1). For the underdoped crystal, $YBa_2Cu_3O_{6.60}$, we observe a sharp (0.5 K wide) discontinuity in the resonance frequency, $\Delta f/f \approx 10^{-4}$, at the superconducting transition (Fig. 1). A sharper discontinuity is observed in the overdoped crystal, $YBa_2Cu_3O_{6.98}$, a possible consequence of the decrease in oxygen disorder near optimal doping. The step discontinuity in resonance frequency and the accompanying discontinuous change (break) in slope are thermodynamic signatures of the superconducting transition (Supplementary Information).

RUS measurements across the temperature range encompassing the pseudogap in the two YBCO crystals are shown in Fig. 2. The temperature dependence of the resonance frequencies in underdoped $YBa_2Cu_3O_{6.60}$ reveals a break in slope at the pseudogap boundary $T^* = 245K$—in itself a standard thermodynamic marker for a phase transition (Fig. 2a, c). It differs from the signature of the superconducting transition in that there is no resolvable discontinuity in the frequency itself. This temperature is the same as the onset temperature of magnetic order observed by neutron scattering measurements of YBCO specimens of similar composition (Fig. 3)[8,9]. In the overdoped crystal, $YBa_2Cu_3O_{6.98}$, the break in slope of the temperature dependence is observed at $T^* = 68$ K (Fig. 2b). To emphasize the break in slope in these data, we use the redundant information contained in all observed resonances to extract the different contributions to their temperature dependences (Fig. 4c). This process reduces the temperature dependence of all 15 normal modes measured to three dominant

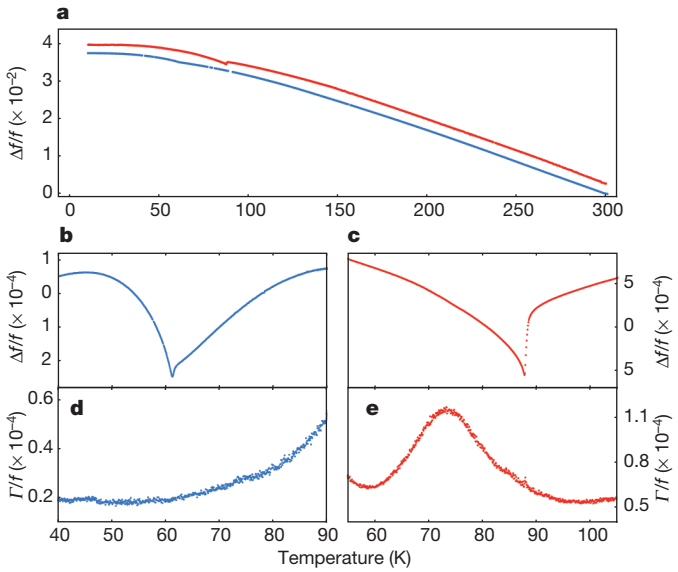


Figure 1 | The temperature evolution of resonances in underdoped and overdoped YBCO crystals: superconductivity. **a**, A typical resonance frequency scan (normalized at room temperature) from room temperature to 10 K for underdoped $YBa_2Cu_3O_{6.60}$ (blue) with $T_c = 61.6\,K$, and overdoped $YBa_2Cu_3O_{6.98}$ (red) with $T_c = 88$ K. The scan for the overdoped crystal is offset vertically for clarity. The smooth increase in frequency, which saturates at low temperature, is driven by the anharmonicity of the lattice and is typical of most solids[29]. **b**, **c**, Superconducting transition in the underdoped (**b**) and overdoped (**c**) crystals. Measurements were made at roughly 70 mK steps. The elastic moduli decrease discontinuously at the transition. The discontinuity is roughly 1 part in 10^{-4} in the underdoped crystal, and 5 parts in 10^{-4} in the overdoped crystal. The form of the smooth monotonic background subtracted to obtain **b** and **c** was chosen only to emphasize the discontinuity[19]. **d**, **e**, Resonance width for underdoped (**d**) and overdoped (**e**) YBCO. In the underdoped crystal no feature at the superconducting transition can be resolved. A broad maximum in resonance width well below T_c in the overdoped crystal is an effect of the pseudogap (see the text).

[1]Pulsed Field Facility, National High Magnetic Field Laboratory, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [2]Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada. [3]Canadian Institute for Advanced Research, Toronto, Canada, M5G 1Z8. [4]Stanford Institute of Materials and Energy Sciences, Stanford University, Stanford, California 94305, USA. [5]Departments of Physics and Applied Physics, and Geballe Laboratory for Advanced Materials, Stanford University, Stanford, California 94305, USA.

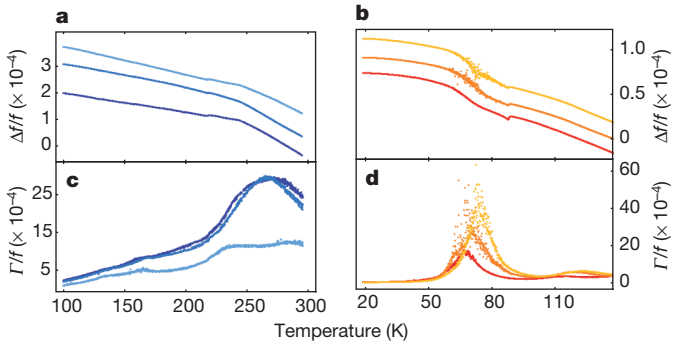


Figure 2 | The temperature evolution of resonances across the pseudogap phase boundary. **a**, **b**, At both dopings a discontinuous change in slope of the temperature dependence of the frequency reveals a phase transition: underdoped (**a**) at $T^* = 245$ K, and overdoped (**b**) at $T^* = 68$ K. **c**, **d**, At both dopings the resonance width has a broad maximum above T^* (underdoped (**c**) and overdoped (**d**)). The break in slope is 5 K wide in the underdoped crystal, and 3 K wide in the overdoped crystal. The increase in scatter of points near the break in slope in **b** is a result of a strong increase in resonance width at this temperature (**d**).

components (see Supplementary Information). The blue and red curves in Fig. 4c capture the effects of superconductivity and of fluctuations in the vicinity of the pseudogap, respectively. The green curve, which has a break in slope at $T^* = 68$ K, corresponds to the thermodynamic effects at the pseudogap, revealing that the pseudogap occurs by means of a phase transition.

The 'strange metal' behaviour that copper oxide superconductors show universally at higher temperature breaks down in the pseudogap region of the temperature–doping phase diagram[2,7,20–24], where measurements indicate the presence of magnetic order[8–10,13] (Fig. 3). The break in slope that we observe in both underdoped and overdoped YBCO establishes the pseudogap as a thermodynamic phase that moves to lower temperature with increased doping. Observation of the pseudogap boundary below the superconducting transition temperature in overdoped YBCO indicates that the superconducting dome surrounds the zero-temperature end point of the pseudogap phase boundary.

At both dopings the pseudogap is accompanied by a large (up to 100-fold in the overdoped crystal) increase in the width of the resonances at temperatures above the pseudogap phase boundary (Fig. 2c, d). The widths of the resonances are determined by the ultrasonic energy absorption (attenuation)[25], revealing strong fluctuations in the dynamics of the metallic state as it approaches T^*. From the width of the resonances we estimate the thermodynamic effects accompanying the pseudogap phase transition to be $\Gamma/f \approx 5 \times 10^{-3}$, about 50-fold the relative modulus shift across the superconducting phase transition for both dopings. Energy absorption is highest when the measurement frequency matches the characteristic relaxation time of the system: $2\pi f\tau(T) = 1$. The characteristic time τ diverges as the phase transition temperature is approached (critical slowing down)[26]; the maximum in ultrasonic energy absorption is therefore closer to the pseudogap phase boundary for resonances of lower frequency. For the underdoped crystal, the width of the maximum and the contribution of the large phonon background at 245 K obscures this effect. The overdoped crystal, with its narrower maxima and smoother background, shows this effect clearly: $1/\tau(T)$ extrapolated from resonances at different frequencies vanishes at the pseudogap phase boundary (Fig. 4a, b). Causality requires that the maxima in energy absorption be accompanied by elastic stiffening over the same temperature range. This stiffening is observed in addition to the distinct break in slope at T^* (Fig. 2b).

The potential for RUS to determine the broken symmetry in the pseudogap phase was limited in this study by the precision with which crystal shape could be controlled, an issue that may be resolvable as sample preparation techniques improve. The pseudogap phase transition

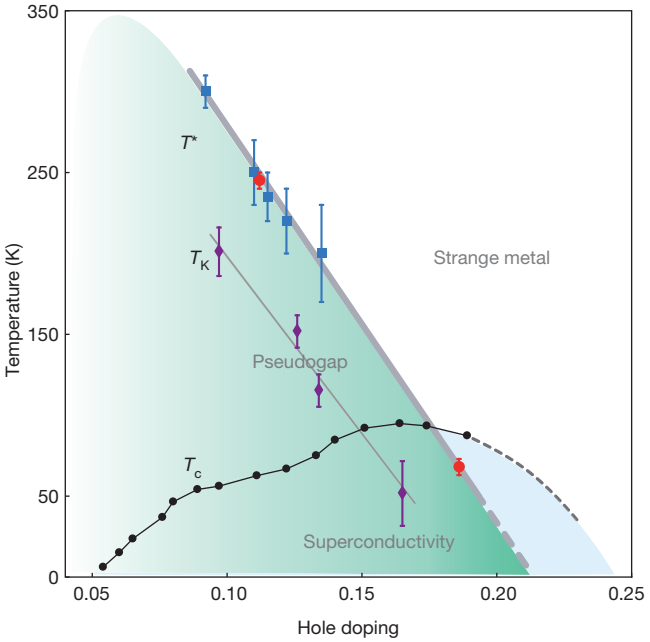


Figure 3 | The phase diagram of $YBa_2Cu_3O_{6+\delta}$. The pseudogap boundary in YBCO is indicated by a thick grey line (guide to the eye), as determined by neutron diffraction measurements[8,9] (blue squares) and resonant ultrasound (red circles). The superconducting transition temperature is indicated by black circles[30]. The temperature of the onset of Kerr rotation[27], where recent X-ray measurements detect an onset of charge order[28], is shown by purple diamonds. Error bars represent the uncertainty in the determination of the onset temperature. The thin grey line is a guide for the eye.

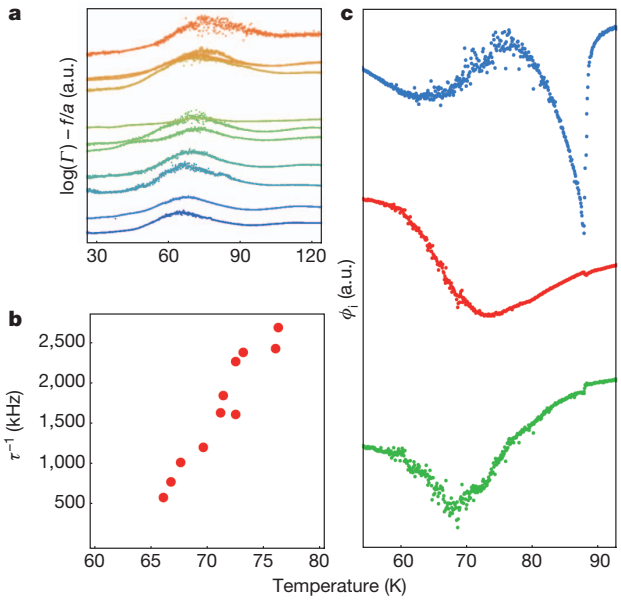


Figure 4 | The pseudogap boundary inside the superconducting dome. **a**, Evolution of resonance width with temperature across T^* for several resonances. To illustrate the evolution of the maximum in resonance width with resonance frequency each curve is offset vertically by an amount proportional to resonance frequency. **b**, Evolution of the temperature of the resonance width maxima in **a** with resonance frequency ($2\pi f\tau = 1$). The characteristic time τ increases as the pseudogap temperature is approached (critical slowing down). **c**, Three different components, $\phi_{i=1,2,3}(T)$, of the temperature dependence of all resonance modes in the overdoped crystal: blue is dominated by superconductivity, red by fluctuations, and green by the pseudogap. The smooth anharmonic background, which dominates Fig. 1a, is not shown. Each curve is scaled vertically for clarity. a.u., arbitrary units.

is located by our RUS measurements with ± 3 K uncertainty, improving on the ± 30 K uncertainty in onset of neutron spin-flip scattering. This clearly separates the onset of magnetic order[8–10,13] at T^* from the onset T_K of the Kerr rotation signal[27] and charge order[28] at lower temperature (Fig. 3). In our measurements we observe an increase in energy absorption over a broad region near T_K (Fig. 2c); however, we do not observe an accompanying thermodynamic signature there. Our observed evolution of the pseudogap phase boundary from underdoped to overdoped establishes the presence of a quantum critical point inside the superconducting dome, suggesting a quantum-critical origin for both the strange metallic behaviour and the mechanism of superconducting pairing.

Received 4 October 2012; accepted 5 April 2013.

1. Ando, Y., Komiya, S., Segawa, K., Ono, S. & Kurita, Y. Electronic phase diagram of high-T_c cuprate superconductors from a mapping of the in-plane resistivity curvature. *Phys. Rev. Lett.* **93,** 267001 (2004).
2. Hussey, N. E. Phenomenology of the normal state in-plane transport properties of high-T_c cuprates. *J. Phys. Condens. Matter* **20,** 123201 (2008).
3. van der Marel, D. *et al.* Quantum critical behaviour in a high-T_c superconductor. *Nature* **425,** 271–274 (2003).
4. Orenstein, J. & Millis, A. J. Advances in the physics of high-temperature superconductivity. *Nature* **288,** 468–474 (2000).
5. Varma, C. M., Littlewood, P. B., Schmitt-Rink, S., Abrahams, E. & Ruckenstein, A. Phenomenology of the normal state of Cu–O high-temperature superconductors. *Phys. Rev. Lett.* **63,** 1996–1999 (1989).
6. Varma, C. M., Nussinov, Z. & van Saarloos, W. Singular or non-Fermi liquids. *Phys. Rep.* **361,** 267417 (2002).
7. Timusk, T. & Statt, B. The pseudogap in high-temperature superconductors: an experimental survey. *Rep. Prog. Phys.* **62,** 61–122 (1999).
8. Fauqué, B. *et al.* Magnetic order in the pseudogap phase of high-T_c superconductors. *Phys. Rev. Lett.* **96,** 197001 (2006).
9. Mook, H. A., Sidis, Y., Fauqué, B., Baldent, V. & Bourges, P. Observation of magnetic order in a superconducting $YBa_2Cu_3O_{6.6}$ single crystal using polarized neutron scattering. *Phys. Rev. B* **78,** 020506 (2008).
10. Kaminski, A. *et al.* Spontaneous breaking of time-reversal symmetry in the pseudogap state of a high-T_c superconductor. *Nature* **416,** 610–613 (2002).
11. Varma, C. M. Non-Fermi-liquid states and pairing instability of a general model of copper oxide metals. *Phys. Rev. B* **55,** 14554–14580 (1997).
12. Aji, V. & Varma, C. M. Quantum criticality in dissipative quantum two-dimensional *XY* and Ashkin–Teller models: application to the cuprates. *Phys. Rev. B* **79,** 184501 (2009).
13. Li, Y. *et al.* Unusual magnetic order in the pseudogap region of the superconductor $HgBa_2CuO_{4+\delta}$. *Nature* **455,** 372–375 (2008).
14. Migliori, A. & Sarrao, J. M. *Resonant Ultrasound Spectroscopy* (Wiley-Interscience, 1997).
15. Migliori, A. & Maynard, J. D. Implementation of a modern resonant ultrasound spectroscopy system for the measurement of the elastic moduli of small solid specimens. *Rev. Sci. Instrum.* **76,** 121301–121308 (2005).
16. Birss, R. R. *Symmetry and Magnetism* (Wiley-Interscience Inc., 1964).
17. Lei, M. *et al.* Elastic constants of a monocrystal of superconducting $YBa_2Cu_3O_{7-\delta}$. *Phys. Rev. B* **47,** 6154–6156 (1993).
18. Doiron-Leyraud, N. *et al.* Quantum oscillations and the Fermi surface in an underdoped high-T_c superconductor. *Nature* **447,** 565–568 (2007).
19. Bishop, D. J. *et al.* Bulk-modulus anomalies at the superconducting transition of single-phase $YBa_2Cu_3O_7$. *Phys. Rev. B* **36,** 2408–2410 (1987).
20. Walstedt, R. E. *et al.* ^{63}Cu NMR shift and linewidth anomalies in the $T_c = 60$ K phase of YBaCuO. *Phys. Rev. B* **41,** 9574–9577 (1990).
21. Vishik, I. M. *et al.* ARPES studies of cuprate Fermiology: superconductivity, pseudogap and quasiparticle dynamics. *New J. Phys.* **12,** 105008 (2010).
22. Daou, R. *et al.* Broken rotational symmetry in the pseudogap phase of a high-T_c superconductor. *Nature* **463,** 519–522 (2010).
23. Kondo, T. *et al.* Disentangling Cooper-pair formation above the transition temperature from the pseudogap state in the cuprates. *Nature Phys.* **7,** 21–25 (2011).
24. Leridon, B., Monod, P. & Colson, D. Thermodynamic signature of a phase transition in the pseudogap phase of $YBa_2Cu_3O_x$ high-T_c superconductor. *Europhys. Lett.* **87,** 17011 (2009).
25. Bhatia, A. B. *Ultrasonic Absorption* (Clarendon Press, 1967).
26. Landau, L. D. & Khalatnikov, I. M. On the anomalous absorption of a sound near to points of phase transition of the second kind. *Dokl. Akad. Nauk SSSR* **96,** 469–472 (1954).
27. Xia, J. *et al.* Polar Kerr-effect measurements of the high-temperature $YBa_2Cu_3O_{6+x}$ superconductor: evidence for broken symmetry near the pseudogap temperature. *Phys. Rev. Lett.* **100,** 127002 (2008).
28. Chang, J. *et al.* Direct observation of competition between superconductivity and charge density wave order in $YBa_2Cu_3O_{6.67}$. *Nature Phys.* **8,** 871–876 (2012).
29. Varshni, Y. Temperature dependence of the elastic constants. *Phys. Rev. B* **2,** 3952–3958 (1970).
30. Liang, R., Hardy, W. N. & Bonn, D. A. Evaluation of CuO_2 plane hole doping in $YBa_2Cu_3O_{6+x}$ single crystals. *Phys. Rev. B* **73,** 180505 (2006).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. Abrahams, J. Analytis, P. Bourges, A. Finkel'stein, M. Greven, N. Harrison, K. Modic, C. Varma, I. Vishik and G. Yu for critical reading of the manuscript and informative discussions. Work at Los Alamos National Laboratory (LANL) was supported by National Science Foundation grant DMR-0654118, by the US Department of Energy and by the State of Florida. LANL is operated by LANS LLC. Work at the University of British Columbia was supported by the Canadian Institute for Advanced Research and the Natural Science and Engineering Research Council.

Author Contributions A.S., J.B.B., S.C.R., R.D.McD. and A.M. designed the experiment. A.S., J.B.B. and A.M. built the electronic circuits and the RUS probe. A.S. and F.F.B. wrote the software and analysed the results. B.J.R., R.L., W.N.H. and D.A.B. prepared the YBCO crystals. A.S., B.J.R., R.D.McD. and A.M. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.S. (arkady.shekhter@gmail.com).

LETTER

doi:10.1038/nature12136

Second sound and the superfluid fraction in a Fermi gas with resonant interactions

Leonid A. Sidorenkov[1,2], Meng Khoon Tey[1,2], Rudolf Grimm[1,2], Yan-Hua Hou[3], Lev Pitaevskii[3,4] & Sandro Stringari[3]

Superfluidity is a macroscopic quantum phenomenon occurring in systems as diverse as liquid helium and neutron stars. It occurs below a critical temperature[1,2] and leads to peculiar behaviour such as frictionless flow, the formation of quantized vortices and quenching of the moment of inertia. Ultracold atomic gases offer control of interactions and external confinement, providing unique opportunities to explore superfluid phenomena. Many such (finite-temperature) phenomena can be explained in terms of a two-fluid mixture[3,4] comprising a normal component, which behaves like an ordinary fluid, and a superfluid component with zero viscosity and zero entropy. The two-component nature of a superfluid is manifest in 'second sound', an entropy wave in which the superfluid and the non-superfluid components oscillate with opposite phases (as opposed to ordinary 'first sound', where they oscillate in phase). Here we report the observation of second sound in an ultracold Fermi gas with resonant interactions. The speed of second sound depends explicitly on the value of the superfluid fraction[5], a quantity that is sensitive to the spectrum of elementary excitations[6]. Our measurements allow us to extract the temperature dependence of the superfluid fraction, a previously inaccessible quantity that will provide a benchmark for theories of strongly interacting quantum gases.

Second sound was first measured in liquid helium II (superfluid ^{4}He below 2.2 K), which is the archetype for quantum liquids characterized by strong interactions[7]. Landau developed his theory of two-fluid hydrodynamics[4] to describe this system and its peculiar properties. In liquid helium II, second sound can be generated[8] by local time-dependent heating and detected by observing the propagation of the resulting temperature wave. In this original context, second sound is characterized as a wave that propagates at constant pressure (an isobaric oscillation), whereas first sound, by contrast, is a wave that propagates with constant entropy per particle (an adiabatic oscillation), just like sound in ordinary fluids.

The observation of second sound in ultracold atomic quantum gases has been a long-standing goal. In weakly interacting Bose–Einstein condensed gases, second sound behaves quite differently from the case of liquid helium II. In such systems, the superfluid density coincides with the density of the Bose–Einstein condensed component over the experimentally relevant temperature range, and second sound reduces to an oscillation of the condensate, the thermal component remaining essentially at rest. The corresponding temperature dependence of the speed of sound has been measured[9]. In two other experiments[10,11], the relative motion of the condensate and the thermal component was investigated and frequency shifts and damping effects were observed. In contrast to dilute Bose gases, resonantly interacting Fermi gases[12,13] are characterized by effects of strong interactions. In such systems, superfluidity and thermodynamics[14–18] have been subjects of intensive research. Here the normal component behaves in a deeply hydrodynamic way over a wide range of temperatures, and the spatial overlap between the normal and the superfluid components can be very large also in the presence of harmonic trapping. In this situation, Landau's two-fluid theory can be readily applied, which suggests that resonantly interacting Fermi gases should behave similarly to superfluid helium. In particular, second sound should have the character of an entropy wave.

Our system is an ultracold, superfluid sample of fermionic ^{6}Li atoms, prepared in a highly elongated harmonic trapping potential (Methods) by well-established procedures of laser and evaporative cooling[19]. The sample consists of $N = 3.0 \times 10^5$ atoms in a balanced mixture of the two lowest-spin states, and is about 500 μm long and 20 μm wide. It is characterized by the Fermi temperature $T_F^{trap} \approx 0.9$ μK (Methods). A magnetic bias field of 834 G is applied, which tunes the interaction between the two spin components to lie at the centre of a scattering resonance, where the *s*-wave scattering length diverges (this realizes the 'unitarity limit' of interactions[12–14]). The cloud's temperature, T, is determined by analysing axial density profiles[17,19], using the knowledge of the equation of state (EOS) from ref. 18. The relevant temperature range for the present experiments is between $0.11T_F^{trap}$ and $0.15T_F^{trap}$.

Our method for observing sound propagation builds on the classical scheme to detect the propagation of first sound in Bose–Einstein condensates[20], which was also applied to resonantly interacting Fermi gases[21]. The general idea is to prepare the quantum gas in a trap that is highly elongated, to create a local perturbation, and to detect its one-dimensional propagation as a pulse along the long trap axis. In the case of first sound, it is straightforward to create and detect such an excitation, but it is less obvious how to do so for second sound.

To produce a local excitation of the cloud, we use a repulsive potential created by a tightly focused green laser beam (Methods) that perpendicularly intercepts the trapped sample at its centre (Fig. 1a). To excite first sound, we abruptly turn on the repulsive beam. The local reduction in the trapping potential acts on the superfluid and normal components in the same way and creates a small hump in the axial density distribution. To excite second sound, we keep the green beam's power constant during the whole experimental sequence, except for a short power-modulation burst that contains eight sinusoidal oscillations in 4.5 ms (Methods and Supplementary Information). The fast modulation locally drives the system out of equilibrium, and the subsequent relaxation increases entropy and temperature. The duration of the burst is chosen such that the system can establish a local thermal equilibrium on a length scale that covers the transverse cloud size but is much shorter than the axial extension of the cloud. In all cases, we take care that the excitation remains a small perturbation of the whole system, which globally stays in a thermal equilibrium state.

To detect sound propagation, we record the axial density profile, $n_1(z, t)$, for various time delays, t, after the excitation pulse: $n_1(z, t)$ is the number density integrated over the transverse degrees of freedom. To enhance the visibility of the density perturbation, we subtract a background profile, $\bar{n}_1(z)$, obtained by averaging the profiles over all measured delay times. Our signal, $\delta n_1(z,t) = n_1(z,t) - \bar{n}_1(z)$, is finally normalized to the maximum observed density, $n_{1,max}$.

[1]Institut für Quantenoptik und Quanteninformation, Österreichische Akademie der Wissenschaften, 6020 Innsbruck, Austria. [2]Institut für Experimentalphysik und Zentrum für Quantenphysik, Universität Innsbruck, 6020 Innsbruck, Austria. [3]Dipartimento di Fisica, Universitá di Trento and INO-CNR BEC Center, I-38123 Povo, Italy. [4]Kapitza Institute for Physical Problems, Russian Academy of Sciences, 119334 Moscow, Russia.

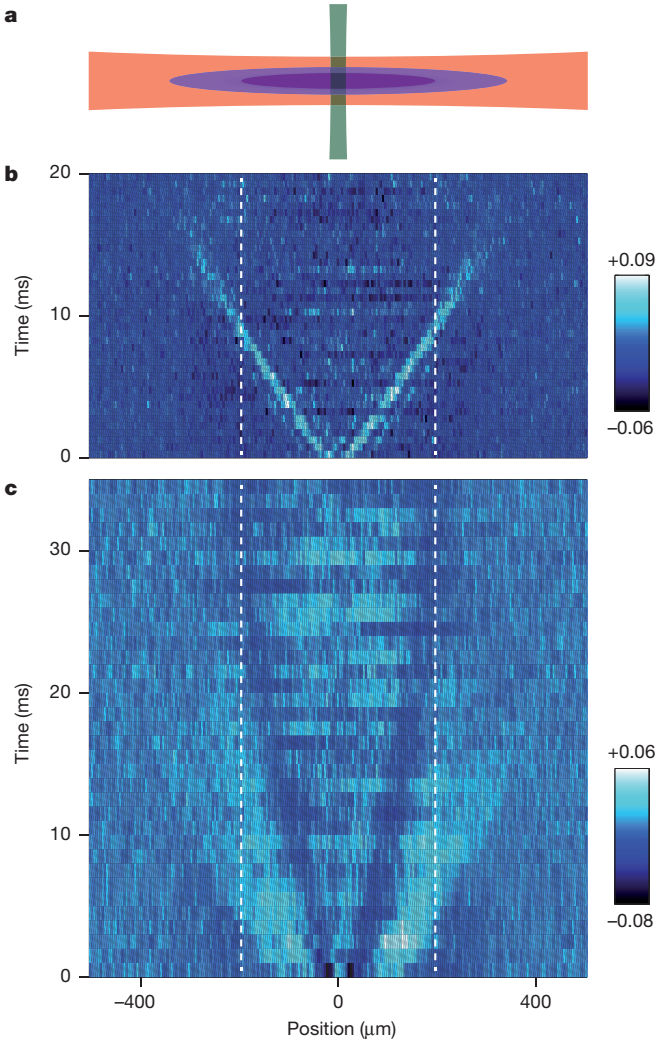


Figure 1 | Observing the propagation of first and second sound. **a**, The basic geometry of exciting the optically trapped cloud with a weak, power-modulated repulsive laser beam (green), which perpendicularly intersects the trapping beam (red). The trapped cloud has a superfluid core ($|z| < 190\,\mu$m), surrounded by a normal region (about 1.5 times larger). **b**, **c**, Normalized differential axial density profiles, $\delta n_1(z,t)/n_{1,\max}$ (colour scale), measured for variable delay times after the excitation show the propagation of first sound (local density increase, bright) and second sound (local decrease, dark). The temperature of the atomic cloud is $T = 0.135(10)\,T_F^{trap}$. The vertical dashed lines indicate the axial region where superfluid is expected to exist according to a recent determination of the critical temperature[18].

The key point for the detection of second sound is the coupling[22,23] between temperature and density variations, which occurs in systems with thermal expansion. The relevant isobaric thermal expansion coefficient can be obtained from the EOS and, for our experimental conditions, is found to be sufficiently large to facilitate the observation of a local temperature increase as a dip in the density profiles (Methods and Supplementary Information).

Evidence of first sound can be seen in Fig. 1b. The initially induced hump splits into two density peaks (bright), which symmetrically propagate outward at an almost constant speed, penetrate into the region where there is no superfluid (outside the dashed lines) and finally fade out in the outer region of the cloud. For longer times, we observe a weak collective breathing oscillation to be excited (not shown).

Excitation using our local heating scheme leads to a strikingly different picture (Fig. 1c). The two density dips (dark) propagate much more slowly than do the first-sound signals. They slow further when approaching the superfluid boundary and finally disappear without penetrating into the non-superfluid region. This behaviour is the telltale characteristic of second sound in our experiment.

To extract the two sound speeds from the differential profiles, $\delta n_1(z,t)$, we determine the positions of the density dips or peaks using Gaussian fitting functions. The corresponding time dependence, extracted from the profiles in Fig. 1, is shown in Fig. 2a. We then use a third-order polynomial to fit globally the time-dependent positions (Fig. 2a, solid lines). The sound speeds are then obtained as time derivatives of the fit curves (Fig. 2b, solid lines). This procedure, by design, produces a smooth curve and does not provide sufficient insight into the uncertainties, so we apply a second procedure to analyse the data on a finer scale. We consider smaller subsets of adjacent points and extract the local speeds by second-order polynomial fits (Supplementary Information). Corresponding results are shown in Fig. 2 by the points.

The fact that the axial harmonic confinement introduces a z dependence of the linear density, n_1, allows us to determine the temperature dependence of the sound speeds, even without changing the global temperature, T, of the trapped sample. The key is to define a z-dependent Fermi temperature, $T_F^{1D} \propto n_1^{2/5}$ (Methods), as the natural local temperature scale. The corresponding reduced temperature, T/T_F^{1D}, has its minimum at the trap centre ($z = 0$) and increases with z. The superfluid phase transition is crossed when the critical temperature, $T_c = 0.214(16)\,T_F^{1D}$, is reached (see ref. 18 for an explanation of the parenthetical uncertainty).

In Fig. 3, we show the temperature dependence of the two speeds of sound, normalized to the local Fermi speed, $v_F^{1D} = \sqrt{2k_B T_F^{1D}/m}$, where m is the atomic mass and k_B is Boltzmann's constant. The symbols correspond to the data displayed in Fig. 2. The solid lines are derived in the same way from the corresponding fit curves. To get additional information on the confidence level of our results, we analysed a number of data sets recorded under similar conditions as the ones in Fig. 1. The regions shaded in pink and light blue show the maximum range of variations considering all our different data sets (Supplementary Information).

Our interpretation of the experimental results relies on an effective one-dimensional (1D) approach to solving Landau's two-fluid hydrodynamic equations for a highly elongated system[24,25]. The basic assumptions are a thermal equilibrium in the radial direction and sufficient shear viscosity to establish a flow field that is independent of the radial position. Within this theoretical framework and under the local density approximation, effective 1D thermodynamic quantities can be defined by integration over the transverse degrees of freedom,

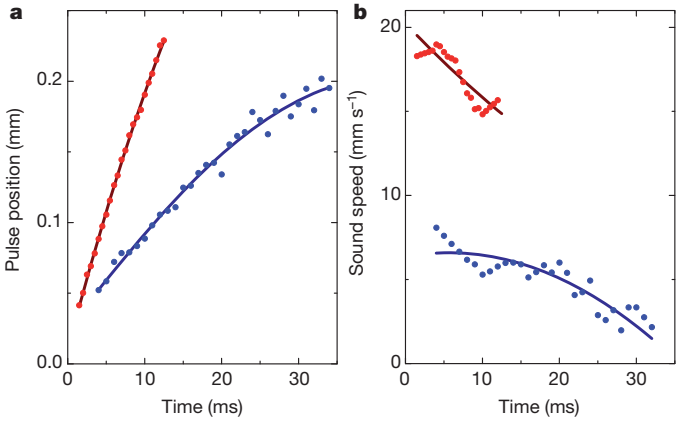


Figure 2 | Extracting the sound speeds. **a**, The positions of the propagating pulses are shown as functions of time. The data points (red and blue symbols for first and second sound, respectively) result from individual fits to the pulses observed at fixed delay times, and the solid lines represent third-order polynomial fits to the time-dependent behaviour. **b**, The sound speeds are obtained as derivatives of the fit curves (solid lines) and, alternatively, by analysing subsets of nine adjacent profiles (points).

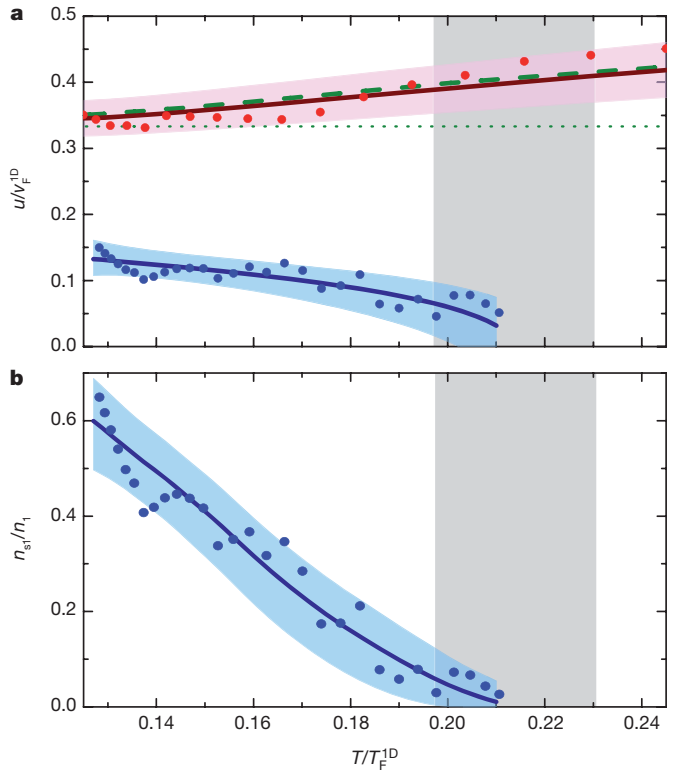


Figure 3 | Normalized sound speeds and the 1D superfluid fraction. **a**, Speeds of first and second sound, normalized to the local Fermi speed and plotted as functions of the reduced temperature. The data points and the solid lines refer to the data set of Fig. 1, following different methods to analyse the raw data (see text). The shaded regions indicate the maximum range of variations from analysing different data sets. The dashed curve is a prediction based on equation (1) and the EOS from ref. 18. The dotted horizontal line is the corresponding zero-temperature limit for the speed of first sound. **b**, Temperature dependence of the 1D superfluid fraction, n_{s1}/n_1, with symbols, solid line and shaded uncertainty range corresponding to the similarly coloured data in **a**. In both panels, the grey shaded area indicates the uncertainty range of the superfluid phase transition according to ref. 18.

such that a thermodynamic quantity, q, yields a 1D counterpart $q_1 \equiv 2\pi \int_0^\infty qr\,\mathrm{d}r$.

At unitarity, we can express the normalized speeds of first and second sound as

$$\frac{u_1}{v_F^{1D}} = \sqrt{\frac{7}{10}\frac{P_1}{n_1 k_B T_F^{1D}}} \qquad (1)$$

and, respectively

$$\frac{u_2}{v_F^{1D}} = \sqrt{\frac{T}{2k_B T_F^{1D}}\frac{\bar{s}_1^2}{\bar{c}_{p1}}\frac{n_{s1}}{n_{n1}}} \qquad (2)$$

where P_1 denotes the 1D pressure (with units of force), $\bar{s}_1 = s_1/n_1$ is the entropy per particle and $\bar{c}_{p1} = T(\partial \bar{s}_1/\partial T)_{p1}$ is the isobaric heat capacity per particle. These thermodynamic quantities can be calculated at unitarity as functions of the reduced temperature, T/T_F^{1D}, using the EOS measured in ref. 18, as we discuss in detail in ref. 25. In contrast, the quantities n_{s1} and $n_{n1} = n_1 - n_{s1}$, which respectively represent the linear number densities of the superfluid and the normal components, cannot be determined from the known EOS.

The speed of first sound provides us with an important benchmark for our experimental method and the interpretation of the measurements in the 1D theoretical framework. The experimental results (Fig. 3a, red symbols and upper solid line) are in excellent agreement with the calculation (dashed line) based on equation (1) and the EOS from ref. 18, which, in addition to recent measurements of the T-dependent frequencies of higher-nodal collective modes[19], is a further confirmation of the validity of our theoretical approach.

The measured speed of second sound (Fig. 3a, blue line and data points) is observed to decrease with increasing temperature, in contrast to first sound. The general behaviour fits very well to qualitative predictions[24]. We can now extract the 1D superfluid fraction, n_{s1}/n_1, which is the unknown quantity in equation (2). The result is presented in Fig. 3b, where n_{s1}/n_1 increases smoothly with decreasing T below the critical temperature.

We finally reconstruct (Methods) the temperature dependence of the superfluid fraction, n_s/n, for the homogeneous three-dimensional case, which has not been measured so far. The results (Fig. 4) turn out to be rather close to those for liquid helium II[26] (solid line). In particular, the gas is almost completely superfluid below $0.6T_c$. This behaviour is quite different from that exhibited by a weakly interacting Bose gas, whose superfluid fraction is significantly smaller, as a function of temperature, than the fraction represented by our data points in Fig. 4 and is well approximated by the condensate fraction of the ideal Bose gas (dashed line). In strongly interacting quantum fluids, the superfluid and the condensate fractions simultaneously appear at the phase transition, but they have quite different temperature dependencies below T_c. Our experimental results provide a benchmark for advancing theoretical approaches to calculate the superfluid fraction, which is a challenging problem in quantum many-body physics (Supplementary Information). With additional tuning of the interaction conditions, corresponding information may be obtained throughout the whole cross-over[12,13,27] from a molecular Bose–Einstein condensate, for which the superfluid fraction should approach the dashed line in Fig. 4, to a Bardeen–Cooper–Schrieffer-type superfluid.

From the point of view of fundamental physics, the experimentally determined superfluid fraction represents a thermodynamic function that has so far been missing and which contains information on the spectrum of elementary excitations and completes the description of the superfluid in terms of universal thermodynamics. From a broader perspective, the creation of second sound is an example of how to control the relative motion of a superfluid with respect to the normal component. This may find applications in various other configurations where resonantly interacting Fermi gases are used as model systems for exploring dynamical and transport phenomena[28–30].

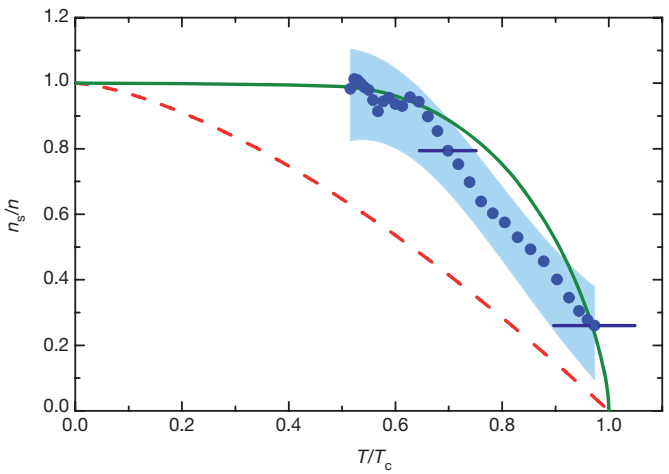


Figure 4 | Superfluid fraction for the homogeneous case. The data points and the corresponding uncertainty range (shaded region) show the superfluid fraction for a uniform, resonantly interacting Fermi gas, reconstructed from its 1D counterpart in Fig. 3b (Methods), as a function of T/T_c. The two horizontal error bars indicate the systematic uncertainties resulting from the limited knowledge of the critical temperature T_c. For comparison, we show the fraction for helium II[26] (solid line) and the textbook expression $1 - (T/T_c)^{3/2}$ for the Bose–Einstein condensed fraction of the ideal Bose gas (dashed line).

METHODS SUMMARY

Our hybrid optical and magnetic trap has radial and axial trapping frequencies of $\omega_r/2\pi = 539(2)$ Hz and $\omega_z/2\pi = 22.46(7)$ Hz, respectively. We use different definitions for the Fermi temperature depending on the trapping geometry under discussion. For a uniform gas with a number density n (including both spin states), the Fermi energy is given by $k_B T_F = (\hbar^2/2m)(3\pi^2 n)^{2/3}$, where $\hbar$ is Planck's constant divided by 2π. For N atoms in a three-dimensional harmonic potential, the corresponding Fermi energy is $k_B T_F^{\mathrm{trap}} = \hbar\left(3N\omega_r^2\omega_z\right)^{1/3}$. Finally, for a cylindrically confined gas where the radial trapping is harmonic, the Fermi temperature is related to the 1D density, n_1, by

$$k_B T_F^{1D} = \left(\frac{15\pi}{8}\right)^{2/5} (\hbar\omega_r)^{4/5} \left(\frac{\hbar^2 n_1^2}{2m}\right)^{1/5}$$

We derive the superfluid fraction, n_s/n, for the uniform case from its 1D counterpart, n_{s1}/n_1, by using the universal thermodynamic relations for a interacting Fermi gas at unitarity. At a given temperature T, the number densities n and n_s can be expressed[14] as $n(x,T) = \lambda_T^{-3} f_n(x)$ and $n_s(x,T) = \lambda_T^{-3} f_{n_s}(x)$, where $\lambda_T = \left(2\pi\hbar^2/mk_B T\right)^{1/2}$ is the thermal de Broglie wavelength. Here the dimensionless parameter $x = \mu/k_B T$, where μ is the chemical potential, is related[25] to T/T_F. The universal function, $f_n(x)$, can be determined from the recent measurements of the EOS[15–18], and the present work allows the function $f_{n_s}(x)$ to be determined. Under the local density approximation, the corresponding 1D densities of a harmonically trapped gas are given[17] by

$$n_1 = \frac{2\pi}{m\omega_r^2}\frac{k_B T}{\lambda_T^3}\int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x$$

and

$$n_{s1} = \frac{2\pi}{m\omega_r^2}\frac{k_B T}{\lambda_T^3}\int_{-\infty}^{x_0} f_{n_s}(x)\,\mathrm{d}x$$

where x_0 is the on-axis value of x. Using these relations, we find that the 1D superfluid fraction, $n_{s1}/n_1 = \int_{-\infty}^{x_0} f_{n_s}(x)\,\mathrm{d}x \Big/ \int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x$, is a universal function depending only on x_0. The uniform superfluid fraction then follows as

$$\frac{n_s}{n} = \frac{f_{n_s}(x_0)}{f_n(x_0)} = \frac{1}{f_n(x_0)}\frac{\mathrm{d}}{\mathrm{d}x_0}\left[\frac{n_{s1}}{n_1}\int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x\right]$$

Full Methods and any associated references are available in the online version of the paper.

Received 9 February; accepted 26 March 2013.
Published online 15 May 2013.

1. Kapitza, P. Viscosity of liquid helium below the λ-point. *Nature* **141,** 74 (1938).
2. Allen, J. F. & Misener, A. D. Flow phenomena in liquid helium II. *Nature* **142,** 643–644 (1938).
3. Tisza, L. Transport phenomena in helium II. *Nature* **141,** 913 (1938).
4. Landau, L. The theory of superfluidity of helium II. *J. Phys. (Mosc.)* **5,** 71–90 (1941).
5. Khalatnikov, I. M. *An Introduction to the Theory of Superfluidity* (Benjamin, 1965).
6. Landau, L. On the theory of superfluidity of helium II. *J. Phys. (Mosc.)* **11,** 91–92 (1947).
7. Atkins, K. R. *Liquid helium* (Cambridge Univ. Press, 1959).
8. Peshkov, V. P. "Second sound" in helium II. *J. Phys. (Mosc.)* **8,** 381 (1944).
9. Meppelink, R., Koller, S. B. & van der Straten, P. Sound propagation in a Bose-Einstein condensate at finite temperatures. *Phys. Rev. A* **80,** 043605 (2009).
10. Stamper-Kurn, D. M., Miesner, H.-J., Inouye, S., Andrews, M. R. & Ketterle, W. Collisionless and hydrodynamic excitations of a Bose-Einstein condensate. *Phys. Rev. Lett.* **81,** 500–503 (1998).
11. Meppelink, R., Koller, S. B., Vogels, J. M., Stoof, H. T. C. & van der Straten, P. Damping of superfluid flow by a thermal cloud. *Phys. Rev. Lett.* **103,** 265301 (2009).
12. Giorgini, S., Pitaevskii, L. P. & Stringari, S. Theory of ultracold atomic Fermi gases. *Rev. Mod. Phys.* **80,** 1215–1274 (2008).
13. Bloch, I., Dalibard, J. & Zwerger, W. Many-body physics with ultracold gases. *Rev. Mod. Phys.* **80,** 885–964 (2008).
14. Ho, T.-L. Universal thermodynamics of degenerate quantum gases in the unitarity limit. *Phys. Rev. Lett.* **92,** 090402 (2004).
15. Kinast, J. *et al.* Heat capacity of a strongly interacting Fermi gas. *Science* **307,** 1296–1299 (2005).
16. Horikoshi, M., Nakajima, S., Ueda, M. & Mukaiyama, T. Measurement of universal thermodynamic functions for a unitary Fermi gas. *Science* **327,** 442–445 (2010).
17. Nascimbène, S., Navon, N., Jiang, K. J., Chevy, F. & Salomon, C. Exploring the thermodynamics of a universal Fermi gas. *Nature* **463,** 1057–1060 (2010).
18. Ku, M. J. H., Sommer, A. T., Cheuk, L. W. & Zwierlein, M. W. Revealing the superfluid lambda transition in the universal thermodynamics of a unitary Fermi gas. *Science* **335,** 563–567 (2012).
19. Tey, M. K. *et al.* Collective modes in a unitary Fermi gas across the superfluid phase transition. *Phys. Rev. Lett.* **110,** 055303 (2013).
20. Andrews, M. R. *et al.* Propagation of sound in a Bose-Einstein condensate. *Phys. Rev. Lett.* **79,** 553–556 (1997).
21. Joseph, J. *et al.* Measurement of sound velocity in a Fermi gas near a Feshbach resonance. *Phys. Rev. Lett.* **98,** 170401 (2007).
22. Arahata, E. & Nikuni, T. Propagation of second sound in a superfluid Fermi gas in the unitary limit. *Phys. Rev. A* **80,** 043613 (2009).
23. Hu, H., Taylor, E., Liu, X.-J., Stringari, S. & Griffin, A. Second sound and the density response function in uniform superfluid atomic gases. *N. J. Phys.* **12,** 043040 (2010).
24. Bertaina, G., Pitaevskii, L. & Stringari, S. First and second sound in cylindrically trapped gases. *Phys. Rev. Lett.* **105,** 150402 (2010).
25. Hou, Y.-H., Pitaevskii, L. & Stringari, S. First and second sound in a highly elongated Fermi gas at unitarity. Preprint at http://arxiv.org/abs/1301.4419 (2013).
26. Dash, J. G. & Taylor, R. D. Hydrodynamics of oscillating disks in viscous fluids: Density and viscosity of normal fluid in pure He^4 from 1.2°K to the lambda point. *Phys. Rev.* **105,** 7–24 (1957).
27. Heiselberg, H. Sound modes at the BCS-BEC crossover. *Phys. Rev. A* **73,** 013607 (2006).
28. Nascimbène, S. *et al.* Collective oscillations of an imbalanced Fermi gas: axial compression modes and polaron effective mass. *Phys. Rev. Lett.* **103,** 170402 (2009).
29. Sommer, A., Ku, M., Roati, G. & Zwierlein, M. W. Universal spin transport in a strongly interacting Fermi gas. *Nature* **472,** 201–204 (2011).
30. Stadler, D., Krinner, S., Meineke, J., Brantut, J.-P. & Esslinger, T. Observing the drop of resistance in the flow of a superfluid Fermi gas. *Nature* **491,** 736–739 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank E. R. Sánchez Guajardo for his contributions in the early stage of this work, and P. van der Straten for discussions. The Innsbruck team acknowledges support from the Austrian Science Fund (FWF) within SFB FoQuS (project no. F4004-N16). The Trento team acknowledges support from the European Research Council through the project QGBE and from the Provincia Autonoma di Trento. We dedicate the present work to our late friend and colleague A. Griffin, who enthusiastically promoted the idea of measuring second sound in Fermi gases.

Author Contributions L.A.S. and M.K.T. equally contributed to the experimental work and the data analysis under the supervision of R.G. The new experimental methods were conceived by these three authors jointly. The theoretical work was performed by Y.-H.H., L.P. and S.S.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.K.T. (mengkhoon.tey@ultracold.at).

METHODS

Trapping potential. In our hybrid trap[31], the tight radial confinement with a trapping frequency $\omega_r/2\pi = 539(2)$ Hz is provided by an infrared laser beam (wavelength, 1,075 nm; power, 120 mW; waist, 39 μm). The much weaker axial confinement, with a trapping frequency $\omega_z/2\pi = 22.46(7)$ Hz, results from the curvature of the applied magnetic field. The frequency ratio, $\omega_r/\omega_z \approx 24$, corresponds to the aspect ratio of the trapped cloud.

Sound excitation and detection. The green laser beam (wavelength, 532 nm; power, <20 mW) used for excitation is focused to a waist in the range of 25–35 μm. To excite first sound, we abruptly turn the laser on to introduce a repulsive potential hill with a height of about 10% of the cloud's chemical potential in the trap centre. For second sound, the repulsive beam is permanently on during the preparation of the quantum gas, with a barrier height of about 15% of the cloud's chemical potential. The excitation is then induced by a burst of eight cycles of sinusoidal power modulation. The modulation frequency is set to 1,720 Hz, but the scheme is quite robust and was found to work best in a frequency range between roughly 2 and 3.5 times the radial trapping frequency. The burst envelope is rectangular and the modulated barrier has a peak-to-peak amplitude of 30% of the cloud's chemical potential. After the burst, the power is set back to its initial constant value. On the timescale of the axial motion, the time-averaged power of the green beam is constant, which avoids direct excitation of first sound.

For the detection of the propagating second-sound signal, the corresponding density dip is essential. The 1D formulation of the EOS in ref. 25 allows us to relate the observed depth to the relative temperature change. In the temperature range of our experiments, the corresponding thermodynamic coefficient, $(\delta n_1/n_1)/(\delta T/T)$, takes values[25] between −0.4 and −0.6, which means that the typical 3% relative depth of the density dip roughly corresponds to a local temperature increase of about 6%.

1D Fermi temperature and critical temperature. Three different definitions for Fermi temperatures are related to natural energy scales of our trapping geometry. The homogeneous case with a 3D number density n (including both spin states) is given by $k_B T_F = (3\pi^2)^{2/3}(\hbar^2/2m)n^{2/3}$. In the local density approximation, the corresponding Fermi energy for N atoms in a 3D harmonic potential is given by $k_B T_F^{\mathrm{trap}} = \hbar\left(3N\omega_r^2\omega_z\right)^{1/3}$, which is commonly used to describe the global situation of a 3D trap. The Fermi energy of the cylindrically confined cloud in the centre of a two-dimensional harmonic trap, which we refer to as '1D Fermi temperature', follows from

$$k_B T_F^{1D} = \left(\frac{15\pi}{8}\right)^{2/5} (\hbar\omega_r)^{4/5} \left(\frac{\hbar^2 n_1^2}{2m}\right)^{1/5}$$

where the relevant density is the linear number density, n_1. For the additional axial confinement in our trap geometry, n_1 and, thus, T_F^{1D} become z-dependent quantities.

To specify the critical temperature, T_c, in units of the relevant Fermi temperature, we also distinguish between the three different situations of a homogeneous system, a 3D harmonic trap and a cylinder with radial harmonic confinement. For the homogeneous system, $T_c = 0.167(13)T_F$ was measured in ref. 18. On the basis of the local density approximation and the experimentally determined EOS, this result can be translated into corresponding conditions for the other two situations. Whereas for the 3D trap, $T_c = 0.223(17)T_F^{\mathrm{trap}}$ is relevant for the occurrence of the phase transition at the centre of the trap, the condition $T_c = 0.214(16)T_F^{1D}$ applies to the radially confined case. In our highly elongated trap geometry with weak axial confinement, T_F^{1D} and, thus, T_c become z dependent. For a fixed global temperature, T, the condition $T < T_c(z)$ then determines the axial range, where a superfluid exists (see illustration in Fig. 1a and dashed lines in Fig. 1b, c).

Reconstruction of superfluid fraction. Within the framework of universal thermodynamics[14], the number density of a uniform, resonantly interacting Fermi gas can be expressed in terms of a dimensionless universal function, $f_n(x)$, as $n(x,T) = \lambda_T^{-3} f_n(x)$. Here $\lambda_T = (2\pi\hbar^2/mk_B T)^{1/2}$ is the thermal de Broglie wavelength, and the dimensionless parameter $x = \mu/k_B T$ gives the ratio between the chemical potential, μ, and the thermal energy, $k_B T$, with a unique correspondence existing between x and T/T_F. The function $f_n(x)$ is known from measurements of the EOS[15–18]. The superfluid density can be expressed as $n_s = \lambda_T^{-3} f_{n_s}(x)$, introducing a corresponding universal function $f_{n_s}(x)$, which is to be extracted from our measurements. Using the local density approximation, one can show for a system with radial harmonic confinement that[17]

$$n_1(x_0,T) = \frac{2\pi}{m\omega_r^2}\frac{k_B T}{\lambda_T^3}\int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x$$

where x_0 represents the value of x on the trap axis. Analogously, the 1D superfluid density is given by

$$n_{s1}(x_0,T) = \frac{2\pi}{m\omega_r^2}\frac{k_B T}{\lambda_T^3}\int_{-\infty}^{x_0} f_{n_s}(x)\,\mathrm{d}x$$

We thus see that the 1D superfluid fraction is given by $n_{s1}/n_1 = \int_{-\infty}^{x_0} f_{n_s}(x)\,\mathrm{d}x \Big/ \int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x$ and only depends on x_0. From our experimental determination of n_{s1}/n_1, we readily obtain the superfluid fraction of a uniform gas using the relation

$$\frac{n_s}{n} = \frac{f_{n_s}(x_0)}{f_n(x_0)} = \frac{1}{f_n(x_0)}\frac{\mathrm{d}}{\mathrm{d}x_0}\left[\frac{n_{s1}}{n_1}\int_{-\infty}^{x_0} f_n(x)\,\mathrm{d}x\right]$$

31. Jochim, S. *et al.* Bose-Einstein condensation of molecules. *Science* **302,** 2101–2103 (2003).

LETTER

doi:10.1038/nature12151

Chemical mapping of a single molecule by plasmon-enhanced Raman scattering

R. Zhang[1]*, Y. Zhang[1]*, Z. C. Dong[1], S. Jiang[1], C. Zhang[1], L. G. Chen[1], L. Zhang[1], Y. Liao[1], J. Aizpurua[2], Y. Luo[1,3], J. L. Yang[1] & J. G. Hou[1]

Visualizing individual molecules with chemical recognition is a longstanding target in catalysis, molecular nanotechnology and biotechnology. Molecular vibrations provide a valuable 'fingerprint' for such identification. Vibrational spectroscopy based on tip-enhanced Raman scattering allows us to access the spectral signals of molecular species very efficiently via the strong localized plasmonic fields produced at the tip apex[1–11]. However, the best spatial resolution of the tip-enhanced Raman scattering imaging is still limited to 3–15 nanometres[5,12–16], which is not adequate for resolving a single molecule chemically. Here we demonstrate Raman spectral imaging with spatial resolution below one nanometre, resolving the inner structure and surface configuration of a single molecule. This is achieved by spectrally matching the resonance of the nanocavity plasmon to the molecular vibronic transitions, particularly the downward transition responsible for the emission of Raman photons. This matching is made possible by the extremely precise tuning capability provided by scanning tunnelling microscopy. Experimental evidence suggests that the highly confined and broadband nature of the nanocavity plasmon field in the tunnelling gap is essential for ultrahigh-resolution imaging through the generation of an efficient double-resonance enhancement for both Raman excitation and Raman emission. Our technique not only allows for chemical imaging at the single-molecule level, but also offers a new way to study the optical processes and photochemistry of a single molecule.

Chemical identification by optical means down to single-molecule sensitivity is a challenging task and usually requires large enhancements of the local fields acting on the molecule[6,7,17–19]. These field enhancements can typically be achieved by using metallic nanoparticles acting as optical antennas to increase the signals[8,11,17,19]. One of the most efficient optical antennas consists of a metallic tip that localizes and enhances optical fields at the tip apex, as in tip-enhanced Raman scattering (TERS), enabling a combination of spectroscopy and microscopy capabilities[1–8,12–18]. However, the spatial extent of the local plasmonic field (5–10 nm) appears to be a limiting factor for spatial resolution[18,20]. Moreover, conventional TERS usually requires the use of strong incident laser fields that could result in undesired diffusion, desorption and even damage to the molecule, thus affecting the sustainability and stability of Raman spectral mapping.

Here we present an experimental study of plasmon-enhanced Raman imaging of single molecules located at the scanning tunnelling microscopy (STM) nanocavity under ultrahigh vacuum and low temperature

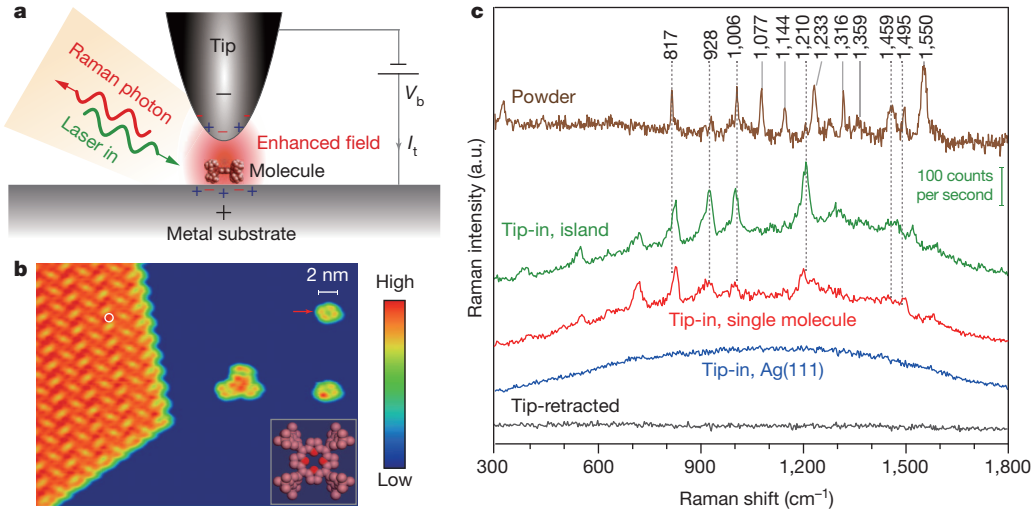


Figure 1 | Clean TERS spectra using well-defined tip and sample. **a**, Schematic tunnelling-controlled TERS in a confocal-type side-illumination configuration, in which V_b is the sample bias and I_t is the tunnelling current. **b**, STM topograph of sub-monolayered H_2TBPP molecules on Ag(111) (1.5 V, 30 pA, 35 nm × 27 nm). The inset shows the chemical structure of H_2TBPP and the white circle indicates one representative site for TERS measurements on the molecular islands. **c**, TERS spectra for different conditions. The tip-in spectra were acquired at 120 mV, 0.5 nA and 3 s. The green spectrum is taken on top of the molecular island (the green scale bar shows the signal level detected by charge-coupled device (CCD). The red spectrum is taken on top of a single molecule (marked by the red arrow in **b**). The blue spectrum is taken on bare Ag(111). The black spectrum is taken on top of the molecular island but with the tip retracted 5 nm from the surface (120 mV, 3 s). For comparison, a standard Raman spectrum (brown) is shown on the top for a powder sample of H_2TBPP molecules.

[1]Hefei National Laboratory for Physical Sciences at the Microscale, University of Science and Technology of China, Hefei, Anhui 230026, China. [2]Material Physics Center CSIC-UPV/EHU and Donostia International Physics Center DIPC, Paseo Manuel de Lardizabal 5, Donostia-San Sebastián 20018, Spain. [3]Theoretical Chemistry and Biology, School of Biotechnology, Royal Institute of Technology, S-10691 Stockholm, Sweden.
*These authors contributed equally to this work.

(Fig. 1a and Supplementary Fig. 1). These special conditions allow us to achieve exquisite tuning between the nanocavity plasmon (or tip plasmon) resonance and the molecular vibronic transitions. In particular, the spectral matching of the nanocavity plasmon resonance to the downward transitions associated with the emission of Raman photons is found to be crucial for enhancing and stabilizing the inelastic scattering signals. Remarkably, the very small incident photon flux used in our STM-controlled spectral-matching TERS enables stable single-molecule mapping with unambiguous chemical identification and unprecedented sub-molecular spatial resolution for a single *meso*-tetrakis(3,5-di-tertiarybutylphenyl)-porphyrin (H_2TBPP) molecule on the Ag(111) surface.

Figure 1b and c illustrate the high quality and level of cleanness of the tip and sample under a low-temperature and ultrahigh-vacuum environment[13,17,21], which allows for reproducible chemical identifications through vibrational fingerprints. In the STM image of Fig. 1b for H_2TBPP on Ag (111), we can identify two isolated single molecules on the right (featuring the characteristic four-lobe pattern[22,23]), a three-molecule cluster at the centre, and a densely packed monolayer island on the left. The bare Ag surface in blue serves as an *in situ* checkground for monitoring the tip cleanness and the nanocavity plasmon resonance mode.

As shown in Fig. 1c, when a tip is positioned on the bare Ag surface, a featureless spectrum (blue) containing simply a broad continuum is obtained, consistent with the absence of molecules and also suggesting the operation of a contamination-free tip. On top of the molecular islands (marked with a white circle in Fig. 1b), the TERS spectrum (green) shows clear vibrational fingerprints of H_2TBPP molecules over a broad continuum. The spectral features do not change substantially when acquired at different positions of molecules on the island (Supplementary Fig. 2). When the tip is retracted about 5 nm from the island (black spectrum), the molecular fingerprints disappear, providing unambiguous evidence that the TERS signals observed in the green spectrum originate only from the molecular sample itself. The far-field Raman signal associated with the tip-retracted mode is absent, so a relatively high signal-to-noise ratio for the TERS signal (for example, the 1,210 cm^{-1} peak) implies a large enhancement in the present system[17,18]. This is nicely illustrated by the TERS measurements on an isolated single H_2TBPP molecule (red spectrum), which exhibits unambiguous vibrational fingerprints similar to those on the molecular island.

We note that inelastic electron tunnelling spectroscopy in cryogenic STM can also provide information about certain vibrational modes of a single molecule through inelastic electron exication[24]. By contrast, the STM-controlled TERS provides full vibrational fingerprints of molecules through the excitation of optical fields alone. Localized tunnelling electrons do not contribute to the TERS signals because the vibrational fingerprints of the H_2TBPP molecule show up in the spectrum even when the bias voltage is well below the excitation threshold of vibrational modes (see Supplementary Fig. 3 and related discussion in the Supplementary Information).

For comparison, the standard Raman signals from a H_2TBPP powder sample are also given in Fig. 1c, showing all Raman-active modes averaged over the full space of randomly oriented molecules. Many of the fingerprint peaks of the powder spectrum are in good correspondence with the vibrational fingerprints on the molecular island and single molecules (see dashed lines), thus providing clear chemical identification of the molecules on the surface. However, the number and relative intensity of peaks reveal differences as well. These could be attributed to the ordering of the H_2TBPP molecules on Ag(111) and the preferred axial polarization of the nanocavity plasmon, which selects particular Raman modes. A detailed spectral assignment is given in the Supplementary Methods and Supplementary Video, obtained using density functional theory calculations.

Of particular interest is the broad continuum that usually accompanies the TERS measurements on both the bare metal surface and the molecules (Fig. 1c)[25–27]. This broad continuum turns out to correlate closely with the nanocavity plasmon resonance mode that is determined by the junction geometry of the STM nanocavity and dielectric properties of the tip and substrate[25], and can be monitored *in situ* by STM-induced luminescence through tunnelling electron excitation[22].

In Fig. 2a, we show two TERS spectra (red) obtained from the 'on-resonance' and 'off-resonance' conditions defined below. 'On-resonance' here refers to good spectral matching between the nanocavity plasmon resonance (blue curves) and the downward molecular vibronic transition of $Q_y(1,1)$. The latter has an energy similar to that of the $Q_y(0,0)$ band measured by the photoluminescence excitation technique (see brown curve at the bottom of Fig. 2a and the schematic of molecular vibronic transitions in Fig. 2b). In contrast, 'off-resonance' refers to the situation in which the nanocavity plasmon resonance does not match well with the $Q_y(1,1)$ band, but does match the laser line. We note that for H_2TBPP molecules, the laser line at 532 nm is always resonant with the upward molecular vibronic $Q_y(1,0)$ transition, but under the 'off-resonance' condition, the TERS spectrum shows very few observable spectral features. In contrast, a dramatic enhancement of the Raman signals can be observed under the 'on-resonance' condition even under a small incident laser flux of about 10^2 W cm^{-2}. The Raman vibrational modes appear as sharp spectral features on top of a smooth continuum, whose profile matches very well with the nanocavity plasmon profile.

In other words, under the 'on-resonance' condition, the broadband nature of the nanocavity plasmon field allows us to realize an efficient

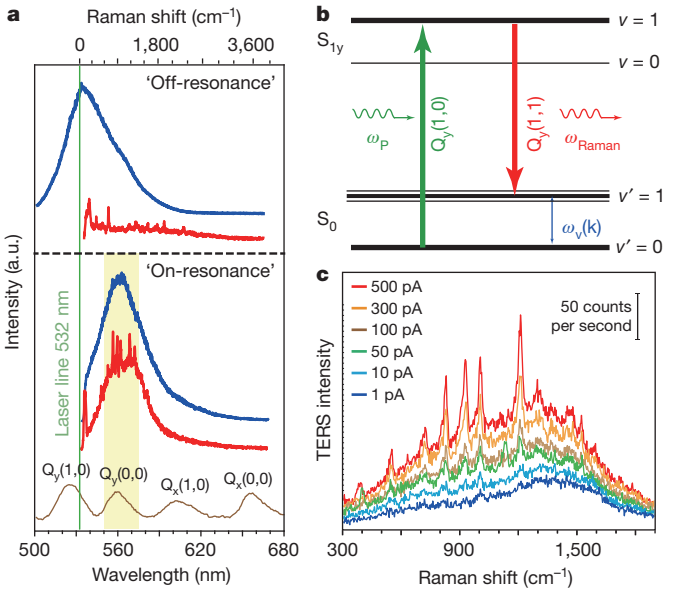


Figure 2 | Spectral matching to generate broadband nanocavity plasmon-enhanced Raman scattering. **a**, Dependence of TERS spectra (red, 200 mV, 1 nA, 3 s) on the spectral matching among the laser line (green), nanocavity plasmon resonance (blue), and molecular vibronic transitions (brown). The lower panel shows the 'on-resonance' situation with the nanocavity plasmon resonance matching the downward molecular vibronic transition $Q_y(0,0)$. The nanocavity plasmon profile (blue) is determined by STM-induced luminescence for the Ag tip on Ag(111) (2.8 V, 0.1 nA, 10 s), whereas the adsorption spectrum of H_2TBPP (brown) is measured by photoluminescence excitation. The upper panel shows the 'off-resonance' situation in which the nanocavity plasmon resonance does not match the downward $Q_y(0,0)$ transition. The scale on the top axis shows the Raman shifts for the TERS measurements. **b**, Schematic of the optical transitions involved in the Raman process between the ground state S_0 and excited state S_1, in which ω_P is the frequency of the incident pumping laser, ω_{Raman} is the frequency of Raman photons, and $\omega_v(k)$ is the vibrational frequency of the k_{th} mode of the molecule; v and v′ stand for the vibrational levels of excited states and ground states, respectively. **c**, TERS spectra on the molecular island as a function of tunnelling currents (120 mV, 3 s).

doubly resonant TERS process: not only producing sufficient resonant excitation through the spectral overlap between the nanocavity plasmon shoulder, the laser line, and the upward molecular vibronic transitions, but, more importantly, also generating large resonant emission enhancement owing to the good spectral matching of the nanocavity plasmon resonance to the downward molecular vibronic transitions. It is worth mentioning that the photon flux used for the 'on-resonance' TERS experiments is about one to two orders of magnitude smaller than what has been reported to date ($\sim 10^3 - 10^4$ W cm^{-2})[6,14,17,21], thus guaranteeing the sustainability of the single molecule during the measurements, that is, it will not be damaged.

Moreover, the broad continuum and associated TERS signals are found to be enhanced dramatically when tunnelling currents are increased (Fig. 2c)[17], which suggests a highly sensitive dependency of both signals on the gap distance (that is, on the local field enhancement) (Supplementary Fig. 4c)[25]. In our spectral-matching TERS, the nanocavity plasmon resonance mode is tuned mainly by modifying the tip status[22] while the strength of the nanocavity plasmon mode is controlled by the gap distance set by the tunnelling condition, particularly the tunnelling current. A sufficiently intense nanocavity plasmon field associated with a relatively short gap distance is important to yield pronounced TERS signals.

However, as shown in Fig. 2a, the proper setting of the nanocavity plasmon resonance to satisfy the 'on-resonance' condition is much more critical to achieve large enhancement and high signal levels. These observations correlate directly with how the nanocavity plasmon field is involved in the Raman scattering process. The overall spectral profile of the 'on-resonance' TERS spectrum in Fig. 2a strikingly resembles that in the broadband femtosecond stimulated Raman scattering process[28]. In addition, we also observed a nonlinear relationship between the TERS response and the incident laser power (see Supplementary Figs 4 and 5). Such nonlinear power dependency, together with the spectral profile similarity to the broadband stimulated Raman scattering, suggests that our spectral-matching TERS can be viewed as an analogue to the third-order nonlinear stimulated Raman scattering process[28], providing both enhanced signals and improved spatial resolution (see Supplementary Information for more details).

The large enhancement thus gained by exploiting the broadband nanocavity-plasmon-stimulated Raman process not only allows for non-invasive TERS measurements with single-molecule sensitivity, but also provides an opportunity to explore the influence of molecular orientations on the vibrational spectral features. Figure 3a shows two representative TERS spectra acquired on two isolated molecules with different adsorption configurations (or molecular orientations). As shown in the STM images on the right (which remain the same after the TERS measurements), one spectrum is acquired on a single flat-lying H_2TBPP molecule adsorbed on the terrace of Ag(111) (blue spectrum), and the other is on a single tilted molecule adsorbed at the step edge (red spectrum). Both spectra reveal vibrational fingerprints characteristic of the H_2TBPP molecules (Fig. 1c) in terms of Raman peak positions.

However, there also exist differences in the relative peak intensities between the two spectra in Fig. 3a, reflecting the changes in the adsorption configurations (or molecular orientations). With the help of the density functional theory calculations under the dipole approximation,

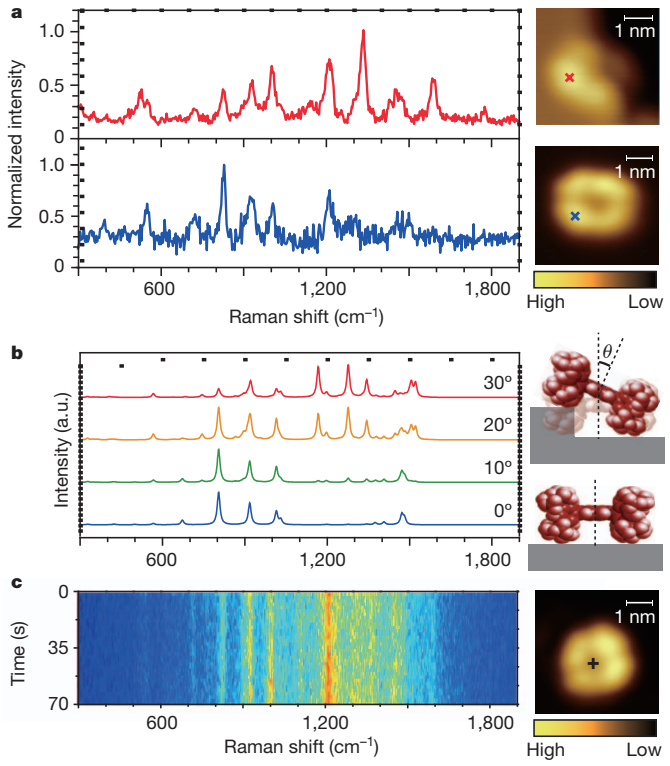


Figure 3 | Single-molecule TERS spectra and their dependency on molecular orientations. **a**, Single-molecule TERS spectra (100 mV, 1 nA, 3 s) for an isolated H_2TBPP molecule adsorbed on the terrace (bottom, blue) or at the step edge (top, red) of Ag(111). Both spectra were acquired on the molecular lobes marked with crosses in the STM images on the right (subtracted from the broad continuum for clarity). **b**, Calculated TERS spectra of a molecule for different tilt angles θ. Shown on the right are the schematics of flat-lying and tilted molecules, respectively. **c**, 35 sequential TERS spectra (120 mV, 1 nA, 2 s) acquired on the centre of a single flat-lying H_2TBPP molecule adsorbed on the Ag(111) terrace (the terrace is the flat area between the step edges); the centre is marked as a cross in the corresponding STM image on the right.

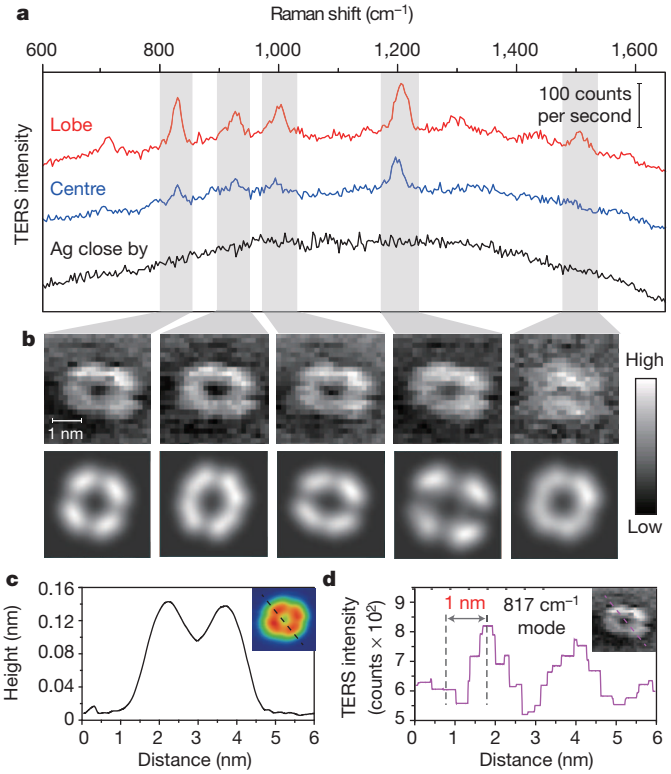


Figure 4 | TERS mapping of a single H_2TBPP molecule on Ag(111). **a**, Representative single-molecule TERS spectra on the lobe (red) and centre (blue) of a flat-lying molecule on Ag(111). The TERS spectrum on the bare Ag about 1 nm away from the molecule is also shown, in black (120 mV, 1 nA, 3 s). **b**, The top panels show experimental TERS mapping of a single molecule for different Raman peaks (23 × 23, ~0.16 nm per pixel), processed from all individual TERS spectra acquired at each pixel (120 mV, 1 nA, 0.3 s; image size: 3.6 × 3.6 nm^2). The bottom panels show the theoretical simulation of the TERS mapping. **c**, Height profile of a line trace in the inset STM topograph (1 V, 20 pA). **d**, TERS intensity profile of the same line trace for the inset Raman map associated with the 817 cm^{-1} Raman peak, integrated over 800−852 cm^{-1}.

the origin of such differences can be associated to the particular orientation of the single molecule that governs the Raman selection rules of the fingerprints. The simulated spectra in Fig. 3b correctly interpret the activation and deactivation of the experimental modes observed in Fig. 3a when the tilting angle of the molecule is changed (see Supplementary Methods for more information). As often found in experiments, similar spectral variations can also occur on the same molecule during the TERS measurements, owing to configurational changes, particularly under the strong illumination condition. Previous single-molecule TERS studies under ambient conditions are usually based on statistical analysis of such spectral fluctuations within a timescale of several seconds[6,7,9,10,17]. However, a single molecule can also remain stable for a certain period of time during the TERS measurements without revealing any substantial spectral changes.

Remarkably, Fig. 3c shows one example of such stability over 70 s when 35 sequential TERS measurements were made on a single H_2TBPP molecule on the Ag terrace. The STM image on the right was taken before the sequential TERS measurements, but remains the same after the measurements. Such molecular stability over a sufficiently long period of time is crucial to perform TERS imaging experiments that can yield a complete and meaningful spectral map.

The realization of efficient nanocavity-plasmon-stimulated Raman scattering in our STM-controlled technique provides a means of performing Raman mapping with unprecedented spatial resolution. Figure 4a indicates that, for an isolated single H_2TBPP molecule on Ag(111), the TERS spectra acquired on the molecular lobe are stronger than those on the centre, whereas the nearby Ag gives a broad continuum, highlighting the highly localized nature of the TERS signals. Such contrast is best illustrated in the panoramic TERS mapping image in Fig. 4b (top row), in which we plot experimental TERS mapping results of a flat-lying molecule on Ag for five selected Raman peaks. The characteristic four-lobe pattern of a H_2TBPP molecule is discernible in the TERS mapping, at least for low-wavenumber vibrational modes. The molecular lobes appear bright but the centre appears dark. However, for relatively large wavenumbers of 1,210 cm^{-1} or above, the contrast and central dark area become smaller.

Such Raman image contrast and frequency dependence can be qualitatively understood by the relative locality of the vibrational modes of the H_2TBPP molecule with respect to the axial polarization of the highly confined local plasmonic fields[6] (Supplementary Video). In brief, the low-wavenumber vibrational modes are more localized in the lobe while the high-wavenumber modes above about 1,210 cm^{-1} contain more contribution from the porphyrin core. To help to explain the contrast between these Raman images and their evolution, the bottom of Fig. 4b shows TERS mapping by assuming that the enhanced local electric field follows a Gaussian beam distribution (Supplementary Methods). The simulated images are consistent with the experimental TERS mapping results, although the consistency is better at low-wavenumber vibrational modes. Strikingly, in comparison with the spatial resolution of the STM topograph (Fig. 4c), the profile of the Raman spectral mapping shown in Fig. 4d not only exhibits a comparable spatial resolution below 1 nm (about 0.5 nm estimated within a 10% to 90% contrast), but more importantly, also provides chemically resolved information revealing intramolecular features.

The ability to access the structure and conformation of a single molecule with both chemical recognition and subnanometre resolution by optical means as presented here provides a new potential to explore the nanometre-scale world, offering new ways to design, control and engineer the functionality of molecules on demand. This should substantially affect the fields of nanophotonics, biochemistry, surface science and molecular electronics, in which identifying molecular species with single-molecule resolution is important. Furthermore, our findings open up a new avenue for studying nonlinear optical processes and photochemistry at the single-molecule scale.

METHODS SUMMARY

Our STM-controlled TERS experiments were performed on a custom low-temperature and ultrahigh-vacuum STM (Unisoku) in a confocal-type side-illumination configuration[29] at about 80 K under a base pressure of around 10^{-10} torr (Supplementary Fig. 1). H_2TBPP molecules were thermally evaporated onto the Ag(111) surface (previously cleaned by argon ion sputtering and annealing). Silver was also used as tip material because the nanogap defined by the Ag tip and the Ag substrate can offer very strong plasmonic resonance[30]. After fabrication via electrochemical etching, Ag tips were cleaned in ultrahigh vacuum via outgassing and ion sputtering, with the tip status further modified by high-voltage pulses[30]. STM imaging and spectral measurements were taken in a constant-current mode with the sample biased. The photon collection and detection systems were described previously[22,30]. A continuous-wave laser at 532 nm is used as a Raman pumping source with a photon flux of about 100 W cm^{-2} illuminating over the junction area.

Received 20 November 2012; accepted 25 March 2013.

1. Stöckle, R. M., Suh, Y. D., Deckert, V. & Zenobi, R. Nanoscale chemical analysis by tip-enhanced Raman spectroscopy. *Chem. Phys. Lett.* **318,** 131–136 (2000).
2. Anderson, M. S. Locally enhanced Raman spectroscopy with an atomic force microscope. *Appl. Phys. Lett.* **76,** 3130–3132 (2000).
3. Hayazawa, N., Inouye, Y., Sekkat, Z. & Kawata, S. Metallized tip amplification of near-field Raman scattering. *Opt. Commun.* **183,** 333–336 (2000).
4. Pettinger, B., Picardi, G., Schuster, R. & Ertl, G. Surface enhanced Raman spectroscopy: towards single molecular spectroscopy. *Electrochem. Jpn.* **68,** 942–949 (2000).
5. Anderson, N., Hartschuh, A., Cronin, S. & Novotny, L. Nanoscale vibrational analysis of single-walled carbon nanotubes. *J. Am. Chem. Soc.* **127,** 2533–2537 (2005).
6. Neacsu, C. C., Dreyer, J., Behr, N. & Raschke, M. B. Scanning-probe Raman spectroscopy with single-molecule sensitivity. *Phys. Rev. B* **73,** 193406 (2006).
7. Sonntag, M. D. *et al.* Single-molecule tip-enhanced Raman spectroscopy. *J. Phys. Chem. C* **116,** 478–483 (2012).
8. Berweger, S. *et al.* Optical nanocrystallography with tip-enhanced phonon Raman spectroscopy. *Nature Nanotechnol.* **4,** 496–499 (2009).
9. van Schrojenstein Lantman, E. M., Deckert-Gaudig, T., Mank, A. J. G., Deckert, V. & Weckhuysen, B. M. Catalytic processes monitored at the nanoscale with tip-enhanced Raman spectroscopy. *Nature Nanotechnol.* **7,** 583–586 (2012).
10. Liu, Z. *et al.* Revealing the molecular structure of single-molecule junctions in different conductance states by fishing-mode tip-enhanced Raman spectroscopy. *Nature Commun.* **2,** 305 (2011).
11. Alonso-González, P. *et al.* Resolving the electromagnetic mechanism of surface-enhanced light scattering at single hot spots. *Nature Commun.* **3,** 684 (2012).
12. Ichimura, T. *et al.* Subnanometric near-field Raman investigation in vicinity of a metallic nanostructure. *Phys. Rev. Lett.* **102,** 186101 (2009).
13. Steidtner, J. & Pettinger, B. Tip-enhanced Raman spectroscopy and microscopy on single dye molecules with 15 nm resolution. *Phys. Rev. Lett.* **100,** 236101 (2008).
14. Stadler, J., Schmid, T. & Zenobi, R. Nanoscale chemical imaging using top-illumination tip-enhanced Raman spectroscopy. *Nano Lett.* **10,** 4514–4520 (2010).
15. Yano, T., Verma, P., Saito, Y., Ichimura, T. & Kawata, S. Pressure-assisted tip-enhanced Raman imaging at a resolution of a few nanometres. *Nature Photon.* **3,** 473–477 (2009).
16. Treffer, R., Lin, X. M., Bailo, E., Deckert-Gaudig, T. & Deckert, V. Distinction of nucleobases—a tip-enhanced Raman approach. *Beilstein J. Nanotechnol* **2,** 628–637 (2011).
17. Pettinger, B., Schambach, P., Villagómez, C. J. & Scott, N. Tip-enhanced Raman spectroscopy: near-fields acting on a few molecules. *Annu. Rev. Phys. Chem.* **63,** 379–399 (2012).
18. Berweger, S. & Raschke, M. B. Signal limitations in tip-enhanced Raman scattering: the challenge to become a routine analytical technique. *Anal. Bioanal. Chem.* **396,** 115–123 (2010).
19. Xu, H. X., Aizpurua, J., Käll, M. & Apell, P. Electromagnetic contributions to single-molecule sensitivity in surface-enhanced Raman scattering. *Phys. Rev. E* **62,** 4318–4324 (2000).
20. Aizpurua, J., Hoffmann, G., Apell, S. P. & Berndt, R. Electromagnetic coupling on an atomic scale. *Phys. Rev. Lett.* **89,** 156803 (2002).
21. Jiang, N. *et al.* Observation of multiple vibrational modes in ultrahigh vacuum tip-enhanced Raman spectroscopy combined with molecular-resolution scanning tunneling microscopy. *Nano Lett.* **12,** 5061–5067 (2012).
22. Dong, Z. C. *et al.* Generation of molecular hot electroluminescence by resonant nanocavity plasmons. *Nature Photon.* **4,** 50–54 (2010).
23. Dong, Z. C. *et al.* Vibrationally resolved fluorescence from organic molecules near metal surfaces in a scanning tunneling microscope. *Phys. Rev. Lett.* **92,** 086801 (2004).
24. Stipe, B. C., Rezaei, M. A. & Ho, W. Single-molecule vibrational spectroscopy and microscopy. *Science* **280,** 1732–1735 (1998).
25. Pettinger, B., Domke, K. F., Zhang, D., Picardi, G. & Schuster, R. Tip-enhanced Raman scattering: influence of the tip-surface geometry on optical resonance and enhancement. *Surf. Sci.* **603,** 1335–1341 (2009).
26. Itoh, T. *et al.* Surface-enhanced resonance Raman scattering and background light emission coupled with plasmon of single Ag nanoaggregates. *J. Chem. Phys.* **124,** 134708 (2006).

27. Yorulmaz, M., Khatua, S., Zijlstra, P., Gaiduk, A. & Orrit, M. Luminescence quantum yield of single gold nanorods. *Nano Lett.* **12,** 4385–4391 (2012).
28. Kukura, P., McCamant, D. W. & Mathies, R. A. Femtosecond stimulated Raman spectroscopy. *Annu. Rev. Phys. Chem.* **58,** 461–488 (2007).
29. Wang, X. *et al.* Tip-enhanced Raman spectroscopy for investigating adsorbed species on a single-crystal surface using electrochemically prepared Au tips. *Appl. Phys. Lett.* **91,** 101105 (2007).
30. Zhang, C. *et al.* Fabrication of silver tips for scanning tunneling microscope induced luminescence. *Rev. Sci. Instrum.* **82,** 083101 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Wang, B. Ren, H. X. Xu, Z. Liu, and X. M. Yang for discussions and Unisoku Company for technical assistance. This work is supported by the National Basic Research Program of China, the Strategic Priority Research Program of the Chinese Academy of Sciences, the Natural Science Foundation of China and the Basque Government Project of Excellence (ETORTEK).

Author Contributions R.Z. and Y.Z. contributed equally to this work. Z.C.D. and J.G.H. supervised the project and designed the experiments. R.Z., Y.Z., S.J., C.Z., L.G.C., L.Z., Y.L. and Z.C.D. performed experiments and analysed data. Y.Z., J.A., Y.L., J.L.Y. and Z.C.D. contributed to the interpretation of the data and theoretical simulations. Z.C.D., Y.Z., Y.L., J.A. and J.G.H. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.C.D. (zcdong@ustc.edu.cn) and J.G.H. (jghou@ustc.edu.cn).

LETTER

doi:10.1038/nature12152

Argon isotopic composition of Archaean atmosphere probes early Earth geodynamics

Magali Pujol[1], Bernard Marty[1], Ray Burgess[2], Grenville Turner[2] & Pascal Philippot[3]

Understanding the growth rate of the continental crust through time is a fundamental issue in Earth sciences[1–8]. The isotopic signatures of noble gases in the silicate Earth (mantle, crust) and in the atmosphere afford exceptional insight into the evolution through time of these geochemical reservoirs[9]. However, no data for the compositions of these reservoirs exists for the distant past, and temporal exchange rates between Earth's interior and its surface are severely under-constrained owing to a lack of samples preserving the original signature of the atmosphere at the time of their formation. Here, we report the analysis of argon in Archaean (3.5-billion-year-old) hydrothermal quartz. Noble gases are hosted in primary fluid inclusions containing a mixture of Archaean freshwater and hydrothermal fluid. Our analysis reveals Archaean atmospheric argon with a $^{40}Ar/^{36}Ar$ value of 143 ± 24, lower than the present-day value of 298.6 (for which ^{40}Ar has been produced by the radioactive decay of the potassium isotope ^{40}K, with a half-life of 1.25 billion years; ^{36}Ar is primordial in origin). This ratio is consistent with an early development of the felsic crust, which might have had an important role in climate variability during the first half of Earth's history.

The continents formed by extraction of incompatible elements from the mantle such as those producing radiogenic heat (U, Th, ^{40}K). The extracted elements have been stored at the Earth's surface because the crust is buoyant, that is, less dense than the underlying mantle. Consequently, the development of the continents affected the composition of the mantle and also shaped the thermal regime of the silicate Earth. Yet no consensus exists on the mode of formation and on the growth rate of the crust. Geological units formed during the first billion years are scarce, and the geochemical methods available to model crustal evolution, such as Sm–Nd in shales[7], and U–Pb and Hf isotopes in zircons[1,8], may have difficulty in distinguishing between the reworking of already existing crust and the creation of juvenile crust (although a combination of isotope tracers seems to provide better constraints[1]).

The terrestrial atmosphere has evolved as a result of volatile exchange between the mantle and the surface of our planet. The inert gases in the atmosphere have accumulated for eons and maintain an integral 'memory' of the degassing of the mantle and the crust. Argon isotopes are potentially useful tracers of these exchanges[9]: ^{36}Ar is primordial, and has been thoroughly degassed from the mantle early in Earth's history, whereas ^{40}Ar, present in negligible amounts at the time of terrestrial accretion, has been produced by the decay of ^{40}K. At present ^{40}Ar is the most abundant argon isotope in the atmosphere (the atmospheric ratio of $^{40}Ar/^{36}Ar = 298.6$; ref. 10), a robust indication of terrestrial degassing through time. The atmosphere contains 1.65×10^{18} moles of ^{40}Ar (ref. 11), which corresponds to about half of the total ^{40}Ar produced in the solid Earth (4.0×10^{18} moles of ^{40}Ar for a silicate Earth K content of 280 parts per million, p.p.m.; ref. 12). The mantle has been evolving through convection and partial melting, during which argon was degassed from mantle-derived magmas into the hydrosphere and atmosphere, while K was concentrated into magmas owing to its incompatible nature. This process resulted in the extraction and storage of a fraction of potassium in the growing continental crust. As soon as the continental crust was formed, even partially, the produced radiogenic ^{40}Ar was less easily degassed into the atmosphere. Consequently, the atmospheric $^{40}Ar/^{36}Ar$ ratio has the potential to trace not only mantle activity but also the growth of the continental crust[13], and to constrain the numerous models of mantle–atmosphere evolution that have been proposed[14–17]. Unfortunately, the record of ancient atmospheric argon isotope ratios in sedimentary rocks is severely compromised by subsequent *in situ* ^{40}Ar production as well as by interaction with crustal fluids containing ^{40}Ar from fluid–rock interactions. Only two attempts to measure ancient atmosphere in a single sedimentary rock appear to have been unaffected by the presence of excess ^{40}Ar. Cadogan[18] and Rice *et al.*[19] proposed that the $^{40}Ar/^{36}Ar$ ratio of the atmosphere in the 395-million-year (Myr)-old Rhynie chert in north-eastern Scotland was 294.1 ± 1.5 (re-normalized to a present-day value of 298.56 ± 0.31; ref. 10). This temporal change requires a ^{40}Ar flux of $(6.2 \pm 2.1) \times 10^7$ moles per year from the solid Earth (crust plus mantle) to the atmosphere averaged over the last 400 Myr, which is consistent with a contemporaneous ^{40}Ar flux of $(11 \pm 1) \times 10^7$ moles per year estimated from measurements of atmospheric argon trapped in Antarctic ice over a time period of 780,000 years (ref. 20).

Our sample comes from the 3.5-billion-year (Gyr)-old Dresser Formation (Warrawoona Group, Pilbara Craton) at North Pole, Western Australia. This formation comprises metabasalts and metakomatiites interleaved with three beds of cherty metasediments that have experienced low-grade metamorphism[21]. The lowermost unit is intercalated with several barite beds and is overlain by silicified carbonate. Undeformed pillow basalts are found above the contact with the chert–barite horizon. Some of the pillow basalts host isolated quartz–carbonate aggregates, forming pods. Our sample is from one of these pods, which represent typical mineralization associated with passive hydrothermal circulation of water through shallow crust. Intrapillow quartz crystals contain abundant, 1–25 μm, two-phase (liquid and <5% vapour) aqueous inclusions, that have been extensively studied for their chemistry[22]. Fluid inclusions are randomly distributed throughout the host quartz, which argues for a primary origin. The absence of crosscutting veins, metamorphic overprint, and deformation features affecting basalt pillows and associated pods indicates negligible fluid remobilization and circulation after deposition and crystallization.

The argon and xenon abundances and isotopic compositions, together with K and Cl contents, were measured by vacuum stepwise crushing, followed by stepwise heating of the powder remaining after crushing, using the extended Ar–Ar method[23] (Supplementary Tables 1 and 2). Using this method, samples were irradiated before analysis with neutrons to transform ^{37}Cl and ^{39}K to ^{38}Ar and ^{39}Ar, respectively, to determine the Cl and K contents at the same extraction steps as ^{36}Ar and ^{40}Ar. Our crushing-step data (Supplementary Table 1) confirm the presence of hydrothermal fluids that were previously identified by X-ray microfluorescence[22]: the Cl/K ratios from crushing experiments vary between 3.6 and 9.4 (Supplementary Fig. 1), within the range 2–48 previously observed[22].

[1]CRPG-CNRS, Université de Lorraine, 15 rue Notre Dame des Pauvres, 54501 Vandoeuvre-lès-Nancy Cedex, France. [2]School of Earth, Atmospheric and Environmental Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK. [3]Institut de Physique du Globe de Paris, Sorbonne-Paris Cité, Université Paris Diderot, CNRS, 1 rue Jussieu, 75238 Paris Cedex 5, France.

The $^{40}Ar/^{36}Ar$ and $Cl/^{36}Ar$ ratios clearly correlate (Fig. 1) between a component rich in radiogenic ^{40}Ar and chlorine and having a near-constant $Cl/^{40}Ar$ ratio of 3,245 ± 330 (Supplementary Information), and a second component with low $^{40}Ar/^{36}Ar$ and low $Cl/^{36}Ar$ values. Because potassium was also measured in these extractions, we compute how much ^{40}Ar could have been produced *in situ* ($^{40}Ar_{in\ situ}$) by ^{40}K decay over 3.5 Gyr (Supplementary Table 1). This accumulation can account for only 5% at best of the total ^{40}Ar for the crushing steps, and 25–34% for the heating steps. Thus the correlation of Fig. 1 indicates mixing between a low-$^{40}Ar/^{36}Ar$, low-salinity component that we assume to be water containing dissolved atmospheric gases, and an hydrothermal fluid end-member containing excess ^{40}Ar ($^{40}Ar_{hydrothermal}$), in constant proportion with respect to Cl. The component displaying low Cl contents and low $^{40}Ar/^{36}Ar$ ratios also has low Cl/K ratios (<2; Supplementary Fig. 1), and is most apparent in the stepwise heating release of gases from the crushed samples, possibly preserved in micrometric fluid inclusions (Fig. 1). The low Cl/K ratio cannot be explained by a simple dilution of the hydrothermal fluids released during sample crushing, nor by the occurrence of a seawater component (Cl/K = 57 for modern sea water). It is instead consistent with the occurrence of a palaeo-atmospheric end-member dissolved in freshwater.

An estimate of the atmospheric $^{40}Ar/^{36}Ar$ ratio can be derived from the intercept of the correlation shown in Fig. 1. However, the data must also be corrected for $^{40}Ar_{in\ situ}$. Given that K is measured for all crushing and heating steps, the correction requires only knowledge of the time that argon was trapped in the sample. The Dresser formation is well dated at 3.52 Gyr by the U–Pb method[24], at 3.5 Gyr by the Sm–Nd method[25], and at 3.49 Gyr by the Pb–Pb method[26]. Massive barite from the Dresser formation has a $U–Xe_{fission}$ (where $Xe_{fission}$ represent xenon isotopes produced by the natural fission of ^{238}U) age of 3.7 ± 0.5 Gyr (ref. 27) and contains excesses of ^{130}Xe ($^{130}Xe^*$) from the double-electron capture decay of ^{130}Ba (half-life 6×10^{20} years) in both fluid inclusions and in the matrix, which demonstrate the antiquity of trapped noble gases[27]. Ar–Ar dating of trapped fluids could not be directly determined for the present sample owing to the large contribution of inherited Ar, but in the Methods we present an Ar isotope data analysis that strongly suggests that fluids trapped in the samples are ≥2.7 Gyr old, probably as old as the Dresser unit. Further evidence that hydrothermal quartz can store noble gases over billion-year timescales arises from the study of another hydrothermal quartz sample filling vacuoles in the komatiitic basaltic unit in the Dresser formation. In that sample, *in situ* radiogenic Ar dominates

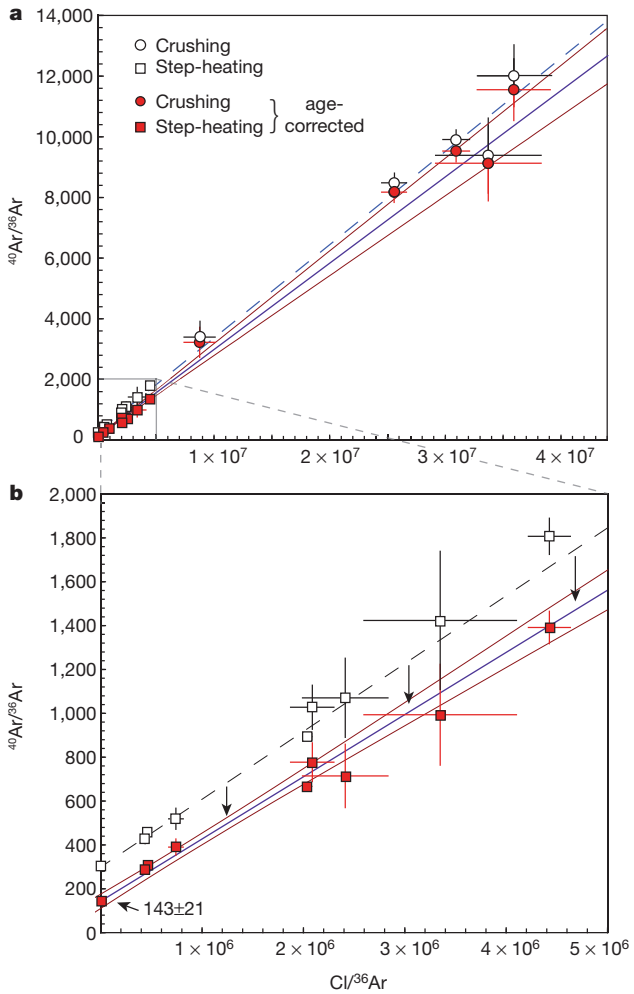


Figure 1 | $^{40}Ar/^{36}Ar$ versus $Cl/^{36}Ar$ for step-heating and step-crushing data of the irradiated sample. **a**, All data. **b**, Enlargement of the stepwise heating data. The data define a two-component mixing trend between an hydrothermal end-member rich in chlorine and inherited $^{40}Ar_{hydrothermal}$, and a low-$^{40}Ar/^{36}Ar$, low-$Cl/^{36}Ar$ end-member representing a low-salinity water component containing dissolved atmospheric gases. The open symbols represent data (from Supplementary Table 1, error bars represent 1σ), and the red symbols represent data age-corrected for *in situ* production of radiogenic ^{40}Ar since the time of fluid trapping. The dotted line and solid line show error-weighted regressions of uncorrected and age-corrected data, respectively). Here an age of 3.5 Gyr has been selected (see Methods).

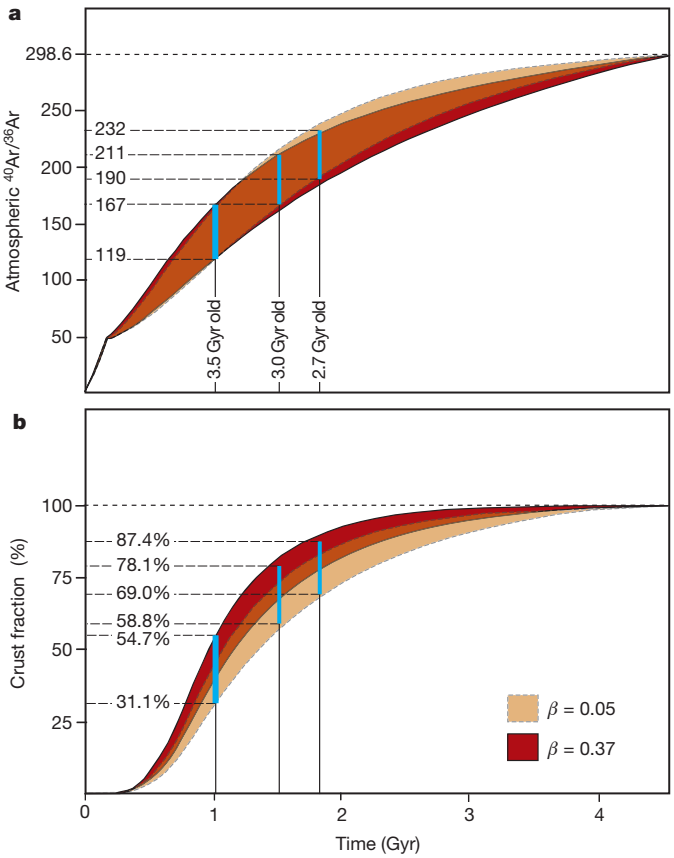


Figure 2 | Evolution of the atmospheric $^{40}Ar/^{36}Ar$ ratio and of the volume of continental crust relative to its present-day volume, as a function of time. **a**, Atmospheric $^{40}Ar/^{36}Ar$ versus time, obtained using our box model (Methods). The shaded areas integrate the trajectories of atmospheric $^{40}Ar/^{36}Ar$ ratio through time for the two extreme rates of crustal degassing ($\beta = 0.05$ for less than 1% crust degassing rate and $\beta = 0.37$ for 50% crust degassing[9]). **b**, Crust fraction versus time. The shaded areas integrate the model runs that fit the conditions defined above. Note that the different boundary conditions we tested (ages of 3.5 Gyr, 3.0 Gyr and 2.7 Gyr, leading to initial $^{40}Ar/^{36}Ar$ ratios of 143 ± 24, 189 ± 21 and 211 ± 21, respectively) yield essentially the same evolution curve for crustal growth.

over the hydrothermal and atmospheric components[28], and yields a Ar–Ar plateau age of 3.0 ± 0.2 Gyr (ref. 28). Both that sample[28] and the one we studied here (Supplementary Information) have radiogenic $^{130}Xe^*$ from the decay of very long-lived ^{130}Ba, and the stable isotope composition of trapped xenon appears to be fractionated (that is, enriched in the light isotopes compared to modern atmospheric Xe; Supplementary Table 2 and Supplementary Fig. 3), a signature of palaeo-atmospheric xenon from the Archaean eon[28]. The last regional metamorphic event took place 2.7 Gyr ago, after which the terranes have been thermally and tectonically stable up to the present[21]. These different lines of evidence, including the textural ones presented above for a primary origin of fluid inclusions, indicate an Archaean age for fluids trapped in this sample, consistent with its formation 3.5 Gyr ago, with a possible lower limit of 2.7 Gyr ago.

After correction for radiogenic ^{40}Ar, the intercept of the mixing correlation yields an Archaean atmospheric $^{40}Ar/^{36}Ar$ of 143 ± 24 (95% confidence interval and mean square weighted deviation, MSWD = 1.5) for age $t = 3.5$ Gyr, obtained using an error-weighted York's regression[29]. Assuming younger fluid ages of 3.0 Gyr and 2.7 Gyr, the initial $^{40}Ar/^{36}Ar$ values are 189 ± 21 and 211 ± 21, respectively. The first heating step at 400 °C released argon with $^{40}Ar/^{36}Ar = 305 \pm 13$, which is consistent with the modern atmospheric $^{40}Ar/^{36}Ar$ (298.6 ± 0.3; ref. 10), and could indicate modern atmospheric contamination. Although during this step ^{39}Ar from neutron irradiation of ^{39}K was also released, suggesting that trapped argon was released at this temperature, we attempted regressions without the 400 °C data, which yielded Archaean atmospheric $^{40}Ar/^{36}Ar$ values of 143 ± 29 (3.5 Gyr), 190 ± 28 (3.0 Gyr) and 212 ± 27 (2.7 Gyr). These values are indistinguishable from those obtained by including the 400 °C step, demonstrating that the results do not depend on the low-temperature step data.

We developed a first-order rate box model following Hamano and Ozima[9], in which the mantle degasses Ar isotopes into the atmosphere through geological time. K is extracted from the mantle during partial melting and most of it is retained in the developing continental crust. The boxes are the mantle, the crust that accumulates K, and the atmosphere. The variables are the mantle extraction rate, the crustal degassing rate for ^{40}Ar (characterized by a β parameter[9] varying between 0.05, representing almost no crustal degassing, and 0.37, corresponding to 50% crustal degassing; see ref. 9 for justification) and the fraction of early degassed ^{36}Ar. The constraints of the model applied to validate the possible solutions are the present-day mantle $^{40}Ar/^{36}Ar$ (5,000 and 40,000 for mantle-plume and mid-ocean-ridge-basalt sources, respectively), the $^{40}Ar/^{36}Ar$ ratio of the modern atmosphere of 298.6 (ref. 10), the palaeo-atmospheric $^{40}Ar/^{36}Ar$ values determined above, the fraction of bulk silicate Earth K in the present-day continental crust (between 20% and 50%; ref. 12), and the mean ^{40}Ar flux to the atmosphere in the last 400 Myr (ref. 18), representing modern conditions. We ran hundreds of tests with this model, allowing us to identify the solutions matching modern conditions and the range of Archaean atmosphere $^{40}Ar/^{36}Ar$ ratios (Supplementary Information). The best solutions indicate (1) catastrophic mantle degassing during the first 170 Myr (impact degassing of accreting bodies cannot be differentiated here); (2) less than 10% stable felsic crust between 170 Myr and 3.8 Gyr; (3) formation of a crustal volume equivalent to 80 ± 10% of the present-day one between 3.8 Gyr and 2.5 Gyr; and (4) less than 30% crustal generation, consistent with possible reworking of previously emplaced felsic crust[1], between 2.5 Gyr and the present day.

The extraction of a large reservoir of felsic crust during the Archaean profoundly modified the thermal regime of the Earth by storing heat-producing radio-elements at the surface. It might have affected the decrease of the partial pressure of atmospheric CO_2, via alteration of this juvenile crust—from the high values of several per cent necessary to prevent Earth's surface from totally freezing when the Sun was about 25% less energetic[30] to the few hundred parts per million that allowed snowball Earth episodes in the late Archaean.

METHODS SUMMARY

We selected quartz because of its generally low content of noble gas-producing elements (for example, K and U). The sample was first neutron-irradiated (to obtain, in addition to natural Ar isotopes, the Cl and K contents), then progressively crushed, and the resulting powder was heated in several temperature steps. ^{36}Ar is predominantly from the atmosphere, but ^{40}Ar can come from three sources, the atmosphere, 'excess' ^{40}Ar from the hydrothermal component ($^{40}Ar_{hydrothermal}$) and ^{40}Ar produced in *situ* ($^{40}Ar_{in\ situ}$) from the decay of ^{40}K. To determine the atmospheric $^{40}Ar/^{36}Ar$ ratio, the measured ^{40}Ar content needs to be corrected for contributions of $^{40}Ar_{hydrothermal}$ and $^{40}Ar_{in\ situ}$. In addition to the geological and geochemical evidence we present, we applied a statistical approach that confirms the Archaean age of the trapped fluids. To do so, we first corrected for the hydrothermal contribution. The hydrothermal $Cl/^{40}Ar_{hydrothermal}$ ratio is obtained from the analysis of the crushing runs, which are dominated (≥95%) by this component (Supplementary Table 3). The step-heating run data are corrected by substracting the step-heating Cl contents multiplied by the $Cl/^{40}Ar_{hydrothermal}$ ratio obtained above. The assumption that the $Cl/^{40}Ar_{hydrothermal}$ ratio of the step-heating and crushing runs are similar is justified by the unique slope of the Fig. 1 correlation. Corrected step-heating data define several equations (one per temperature step) with two unknowns, the amount of *in situ* ^{40}Ar produced, which depends on fluid age, and the initial (atmospheric) $^{40}Ar/^{36}Ar$. We explored the sets of ages and initial $^{40}Ar/^{36}Ar$ values that best fit the equations and found that they correspond to ages around 3.5 Gyr (Methods, Supplementary Table 3 and Supplementary Fig. 2). These ages were then used to correct for $^{40}Ar_{in\ situ}$ in the regression shown in Fig. 1. The initial, presumably palaeo-atmospheric $^{40}Ar/^{36}Ar$ ratio was computed using the error-weighted regression method of York[29].

Full Methods and any associated references are available in the online version of the paper.

Received 25 July 2012; accepted 2 April 2013.

1. Dhuime, B., Hawkesworth, C. J., Cawood, P. A. & Storey, C. D. A change in the geodynamics of continental growth 3 billion years ago. *Science* **335,** 1334–1336 (2012).
2. Hawkesworth, C. J. & Kemp, A. I. S. The differentiation and rates of generation of the continental crust. *Chem. Geol.* **226,** 134–143 (2006).
3. Armstrong, R. L. & Harmon, R. S. Radiogenic isotopes: the case for crustal recycling on a near-steady-state no-continental-growth Earth. *Phil. Trans. R. Soc. Lond. A* **301,** 443–472 (1981).
4. Hurley, P. M. & Rand, J. R. Pre-drift continental nuclei. *Science* **164,** 1229–1242 (1969).
5. McLennan, S. M. & Taylor, R. S. Geochemical constraints on the growth of the continental crust. *J. Geol.* **90,** 347–361 (1982).
6. Reymer, A. & Schubert, G. Phanerozoic addition rates to the continental crust and crustal growth. *Tectonics* **3,** 63–77 (1984).
7. Allègre, C. J. & Rousseau, D. The growth of the continent through geological time studied by Nd isotope analysis of shales. *Earth Planet. Sci. Lett.* **67,** 19–34 (1984).
8. Condie, K. C., Bickford, M. E., Aster, R. C., Belousova, E. & Scholl, D. W. Episodic zircon ages, Hf isotopic composition, and the preservation rate of continental crust. *Geol. Soc. Am. Bull.* **123,** 951–957 (2011).
9. Hamano, Y. & Ozima, M. in *Terrestrial Rare Gases* (eds Alexander, E. C. & Ozima, M.) *Adv. Earth Planet. Sci. Jpn. Sci. Soc.* **3,** 155–171 (1978).
10. Lee, J.-Y. *et al.* A redetermination of the isotopic abundances of atmospheric Ar. *Geochim. Cosmochim. Acta* **70,** 4507–4512 (2006).
11. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* (Cambridge Univ. Press, 2001).
12. Arevalo, R. Jr, McDonough, W. F. & Luong, M. The K/U ratio of the silicate Earth: insights into mantle composition, structure and thermal evolution. *Earth Planet. Sci. Lett.* **278,** 361–369 (2009).
13. Fanale, F. P. A case for catastrophic early degassing of the Earth. *Chem. Geol.* **8,** 79–105 (1971).
14. Pepin, R. O. Atmospheres on the terrestrial planets: clues to origin and evolution. *Earth Planet. Sci. Lett.* **252,** 1–14 (2006).
15. Tolstikhin, I. N. & Marty, B. The evolution of terrestrial volatiles: a view from helium, neon, argon and nitrogen isotope modeling. *Chem. Geol.* **147,** 27–52 (1998).
16. Porcelli, D. & Wasserburg, G. J. Mass transfer of helium, neon, argon, and xenon through a steady-state upper mantle. *Geochim. Cosmochim. Acta* **59,** 4921–4937 (1995).
17. Allègre, C. J., Staudacher, T. & Sarda, P. Rare gas systematics: formation of the atmosphere, evolution and structure of the Earth's mantle. *Earth Planet. Sci. Lett.* **81,** 127–150 (1987).
18. Cadogan, P. H. Paleoatmospheric argon in Rhynie chert. *Nature* **268,** 38–41 (1977).
19. Rice, C. M. *et al.* A Devonian auriferous hot spring system, Rhynie, Scotland. *J. Geol. Soc. Lond.* **152,** 229–250 (1995).
20. Bender, M. L., Barnett, B., Dreyfus, G., Jouzel, J. & Porcelli, D. The contemporary degassing rate of Ar-40 from the solid Earth. *Proc. Natl Acad. Sci. USA* **105,** 8232–8237 (2008).
21. Buick, R. & Dunlop, J. S. R. Evaporitic sediments of early Archaean age from the Warrawoona Group, North Pole, Western Australia. *Sedimentology* **37,** 247–277 (1990).

22. Foriel, J. *et al.* Biological control of Cl/Br and low sulfate concentration in a 3.5-Ga-old seawater from North Pole, Western Australia. *Earth Planet. Sci. Lett.* **228,** 451–463 (2004).
23. Turner, G. Hydrothermal fluids and argon isotopes in quartz veins and cherts. *Geochim. Cosmochim. Acta* **52,** 1443–1448 (1988).
24. Van Kranendonk, M. J., Philippot, P., Lepot, K., Bodorkos, S. & Parajno, F. Geological setting of Earth's oldest fossils in the ca. 3.5 Ga Dresser Formation, Pilbara Craton, Western Australia. *Precambr. Res.* **167,** 93–124 (2008).
25. Tessalina, S. G., Bourdon, B., Van Kranendonk, M. V., Birck, J. L. & Philippot, P. Influence of Hadean crust evident in basalts and cherts from the Pilbara Craton. *Nature Geosci.* **3,** 214–217 (2010).
26. Thorpe, R. I., Hickman, A. H., Davis, D. W., Mortensen, J. K. & Trendall, A. F. U-Pb zircon geochronology of Archaean felsic units in the Marble Bar region, Pilbara Craton, Western Australia. *Precambr. Res.* **56,** 169–189 (1992).
27. Pujol, M., Marty, B., Burnard, P. & Philippot, P. Xenon in Archaean barite: weak decay of ^{130}Ba, mass-dependent isotopic fractionation and implication for barite formation. *Geochim. Cosmochim. Acta* **73,** 6834–6846 (2009).
28. Pujol, M., Marty, B. & Burgess, R. Chondritic-like xenon trapped in Archaean rocks: a possible signature of the ancient atmosphere. *Earth Planet. Sci. Lett.* **308,** 298–306 (2011).
29. York, D. Least-squares fitting of a straight line. *Can. J. Phys.* **44,** 1079–1086 (1966).
30. Kasting, J. F. Faint young Sun redux. *Nature* **464,** 687–689 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Blagburn and L. Zimmermann for their technical support with the irradiated samples measurements, and M. Derrien and B. Faure for their help with the conception of the degassing model. This project was funded by the CNRS, the Région Lorraine, the ANR (Agence Nationale pour la Recherche) projects ''e-Life'' and ''e-Life2'' to P.P. and by the European Research Council under the European Community's Seventh Framework Program (FP7/2007–2013 grant agreement number 267255 to B.M. The drilling programme was supported by funds from the Institut de Physique du Globe de Paris (IPGP) and the CNRS, and by the Geological Survey of Western Australia (GSWA). This is Centre de Recherches Géochimiques et Pétrographiques (CRPG) contribution number 2239.

Author Contributions M.P. and R.B. performed the experiments and analysed the data. P.P. provided the sample and characterized the fluid inclusions. M.P. and B.M. did the calculations and the modelling, and wrote the paper. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.M. (bmarty@crpg.cnrs-nancy.fr).

METHODS

Neutron irradiation and Ar isotope analysis. The argon isotopic analysis of neutron-irradiated quartz (0.09 g), performed at Manchester University in the UK, was used to determine Ar, K (via $^{39}Ar_K$), Cl ($^{38}Ar_{Cl}$), Ca ($^{37}Ar_{Ca}$) concentrations. Neutron irradiation of samples was carried out in position B2W of the SAFARI-1 reactor, NECSA, Pelindaba (South Africa) with a fast neutron fluence of 10^{18} neutrons per cm^2 as determined from Hb3gr flux monitors included in the irradiation. Experimental procedures were similar to those described previously[31]. Samples were progressively crushed *in vacuo* using modified Nupro valves. Liberated gases were purified using hot (400 °C) Al–Zr getters (a getter is a physico-chemical alloy that traps all gases except the noble gases) before being analysed in the mass spectrometer. Samples of crushed residue were step-heated in a tantalum-resistance furnace using several temperature steps, each of 30 min duration. One low-temperature step at 200 °C was used to remove adsorbed atmospheric noble gases from the samples. Sequential temperature steps at 200 °C intervals between 400 °C and 1,600 °C were used to extract argon from the quartz. The argon is likely to be contained in microscopic fluid inclusions, because the siliceous matrix does not contain an appreciable amount of noble gases[23]. For both crushing and stepped heating isotopic measurements were made using the MS1 mass spectrometer using a Faraday detector for Ar isotope measurements. Average furnace hot blanks (1,800 °C) contained 4×10^{-13} moles of ^{40}Ar. Data were corrected for mass discrimination, radioactive decay since irradiation and minor neutron interference corrections obtained from irradiated salts. Concentrations of K, Ca and Cl were determined from samples using Hb3gr monitor data[23].

Fluid composition. Gases released by step crushing reveal the composition of a hydrothermal component having a Cl/K molar ratio between 3.7 and 9.4, with elevated $^{40}Ar/^{36}Ar$ ratio. The range of Cl/K is within hydrothermal end-member values found for large fluid inclusions[22] (Supplementary Fig. 1). Gases released by crushing have lower argon isotope and Cl/K ratios (<2). The $^{40}Ar/^{36}Ar$ and Cl/K values correlate, indicating that the low-$^{40}Ar/^{36}Ar$ component (1) cannot result from dilution of a hydrothermal component by air because of the correlated variation of the Cl/K ratio; and (2) cannot be mixing with sea water (Cl/K = 57 for modern sea water) because the latter would result in an inverse correlation between $^{40}Ar/^{36}Ar$ and Cl/K.

Statistical constraints on the age of trapped fluids. The correlations shown in Fig. 1 use the data given in Supplementary Table 1. To derive the initial $^{40}Ar/^{36}Ar$ ratio, which we propose to be that of the atmosphere at the time of fluid trapping, data need to be corrected for *in situ* production, and the age of trapped fluids is critical. We have given several geological and geochemical arguments that these fluids are Archaean in age, and here we further analyse our data.

Argon-40 ($^{40}Ar_{total}$) trapped in fluid inclusions and in the matrix is a mixture of three components: *in situ* ^{40}Ar ($^{40}Ar_{in\ situ}$) produced since closure of the sample, atmospheric ^{40}Ar ($^{40}Ar_{atmospheric}$) trapped at the time of closure, and inherited argon from the hydrothermal fluid ($^{40}Ar_{hydrothermal}$).

$$^{40}Ar_{total} = {}^{40}Ar_{hydrothermal} + {}^{40}Ar_{in\ situ} + {}^{40}Ar_{atmospheric}$$

In gases extracted by crushing, $^{40}Ar_{total}$ is dominated by $^{40}Ar_{hydrothermal}$ and $^{40}Ar_{atmospheric}$ represents only a few per cent of the total ^{40}Ar. This can be verified for the most extreme conditions, by computing the maximum $^{40}Ar_{in\ situ}$ contribution, assuming both a maximum age of 3.5 Gyr and a $^{40}Ar_{atmospheric}$ content obtained by multiplying the observed ^{36}Ar by the modern value of $(^{40}Ar/^{36}Ar)_{atmospheric}$ (298.6). This is a non-realistic case because it is not possible for both conditions to apply; however, it demonstrates that even in the most extreme case $^{40}Ar_{hydrothermal}$ makes up over 95% of the total ^{40}Ar.

The $Cl/^{40}Ar_{total}$ value of the crushing steps represents the ratio in the hydrothermal fluid end-member at better than 95%. We assume that the $Cl/^{40}Ar$ ratio of the hydrothermal end-member is the same for the crushing runs as for the step-heating runs. The assumption is justified by the fact that the data identify a single hydrothermal end-member having a constant $Cl/^{40}Ar$ ratio, for example, in a $^{40}Ar/^{36}Ar$ versus $Cl/^{36}Ar$ diagram. Thus we correct $^{40}Ar_{total}$ extracted by step-heating for the $^{40}Ar_{hydrothermal}$ contribution, using the mean $(Cl/^{40}Ar_{total})$ value of the crushing runs. In practice, we subtract the measured step-heating Cl ($Cl_{step\text{-}heating}$)from the step-heating $^{40}Ar_{total}$ multiplied by the mean $(Cl/^{40}Ar)_{crushing}$ ratio:

$$(^{40}Ar_{total})_{step\text{-}heating} = (^{40}Ar_{hydrothermal})_{step\text{-}heating} + (^{40}Ar_{in\ situ})_{step\text{-}heating} + (^{40}Ar_{atmospheric})_{step\text{-}heating} \leftrightarrow [(^{40}Ar_{in\ situ})_{step\text{-}heating} + (^{40}Ar_{atmospheric})_{step\text{-}heating}] = (^{40}Ar_{total})_{step\text{-}heating} - Cl_{step\text{-}heating}/(Cl/^{40}Ar_{total})_{crushing}$$

To be independent of the age and of the initial $(^{40}Ar/^{36}Ar)_{atmospheric}$ value, we calculated the $Cl/^{40}Ar_{hydrothermal}$ values of the crushing steps by correcting $^{40}Ar_{total}$ from the (small) contributions of $^{40}Ar_{atmospheric}$ and $^{40}Ar_{in\ situ}$ for ages varying between 0 and 3.5 Gyr and $(^{40}Ar/^{36}Ar)_{atmospheric}$ ratios varying between 100 and 298.6. For all these conditions, the $Cl/^{40}Ar_{hydrothermal}$ value varies between 3,100 and 3,300, which is well within the standard deviation of 330 among the four crushing data (computed with data from Supplementary Table 1). We obtain $Cl/^{40}Ar_{hydrothermal}$ = 3,245 (mean of all these conditions) ± 330 (standard deviation for the four crushing steps). ^{40}Ar from step-heating runs consists now of a mixture of $^{40}Ar_{in\ situ}$ and $^{40}Ar_{atmospheric}$. For each step, we computed the amount of $^{40}Ar_{in\ situ}$ as:

$$(^{40}Ar_{in\ situ})_{step\text{-}heating} = [(^{40}Ar_{in\ situ})_{step\text{-}heating} + (^{40}Ar_{atmospheric})_{step\text{-}heating}] - (^{36}Ar_{atmospheric})_{stepheating} \times (^{40}Ar/^{36}Ar)_{atmospheric}$$

(where $[(^{40}Ar_{in\ situ})_{step\text{-}heating} + (^{40}Ar_{atmospheric})_{step\text{-}heating}]$ has been computed as above). We do not know $(^{40}Ar/^{36}Ar)_{atmospheric}$ a priori, so we consider this ratio as an input parameter for which we assume different values, in practice varying it between 100 and 298.6. With the obtained $(^{40}Ar_{in\ situ})_{step\text{-}heating}$, we compute the corresponding ages as we also have the K concentration for each step.

Thus for each $(^{40}Ar/^{36}Ar)_{atmospheric}$ input value, we obtain a set of step-heating data and we test statistically the homogeneity of ages between the different steps. For that, we computed the Ar–Ar plateau (using the Isoplot software developed by K. Ludwig, http://bgc.org/isoplot_etc/isoplot.html) corresponding to each $(^{40}Ar/^{36}Ar)_{atmospheric}$ value (Supplementary Table 3). The best solutions are those for which ages have the lowest standard deviation and MSWD value closest to 1 (meaning that the errors can account for the spread of data), as given in Supplementary Table 3 and illustrated in Supplementary Fig. 2, and correspond to ages close to the formation age of 3.5 Gyr. For ages lower than 3 Gyr, the MSWD value rapidly becomes close to 0 and the standard deviations increase dramatically. This, together with an age of 3.0 Gyr obtained for a previously analysed quartz sample (for which the *in situ* produced ^{40}Ar was dominant and the hydrothermal contribution was constant for all steps, so that direct Ar–Ar plateau ages could be obtained[27]) as well as with the geological and morphological evidence discussed earlier, points to a palaeo-Archaean age for fluids trapped in quartz, probably the formation age, and excludes a young age for trapped fluids.

Xenon isotopic signature. Xenon isotope analysis was done at CRPG in Nancy, France. Pure quartz grains (1–2 mm in size) were selected and ultrasonically cleaned with acetone. After cleaning, 0.2–0.8 g of the quartz sample was loaded into a stainless steel tube for crushing. The tube was then baked overnight at 150 °C under high vacuum to desorb atmospheric noble gases from the sample surface before extraction. The sample was crushed at room temperature by activating a piston 1,000 times. During crushing, condensable gases including xenon were trapped in a glass cold-finger immersed in liquid nitrogen to separate them from lighter noble gases (He, Ne, Ar). After cryogenic separation, the non-trapped fraction was rapidly pumped, condensable gases were desorbed, and Xe was purified using five successive getters cycled between 700 °C and room temperature. Xe isotopes were then analysed by static mass spectrometry.

The Xe isotope abundances (Supplementary Fig. 3 and Supplementary Table 2), normalized to ^{132}Xe and to the isotopic composition of xenon in modern air, display excesses at masses 126 and 131 (Supplementary Fig. 3a), comparable to excesses reported[32] for an Archaean barite sample, and attributed[32] to cosmic ray spallation reactions forming ^{126}Xe, and production of $^{130}Ba(n,\gamma)^{131}Xe$ by epithermal neutrons[27,32,33]. Interaction with cosmic rays is consistent with the location of the present sample at the surface. Not only are the ^{126}Xe and ^{131}Xe isotopes in excess relative to ^{132}Xe, but also the other lighter Xe isotopes, including ^{130}Xe and ^{129}Xe. ^{130}Xe is itself in excess of ^{129}Xe, indicating the contribution of the natural radioactivity of ^{130}Ba ($^{130}Ba(2EC)^{130}Xe$, (where 2EC indicates double electron capture, a double-decay nuclear reaction) with a half-life of $(6.0 \pm 1.1) \times 10^{20}$ years; ref. 27) and therefore the presence of an old xenon component. Thus the heavy isotopes of xenon ($^{132,134,136}Xe$) must also be contributed by products of the natural fission of ^{238}U and the original Xe isotope composition needs to be corrected.

The U content was measured in these samples (Service d'Analyse des Roches et des Minéraux, CRPG Nancy, France) using two different methods (light leaching of powders to obtain an average U content of fluid inclusions, and U measurement of quartz before any crushing) which both gave a similar U concentration of 0.15 p.p.m. In Supplementary Fig. 3b, the heavy isotope abundances of Xe are corrected for contribution of fissiogenic Xe over 3.5 Gyr (using the younger fluid ages of 3.0 or 2.7 Gyr results in smaller but essentially comparable corrections; Supplementary Fig. 3b). The corrected Xe abundance is clearly deficient in heavy Xe isotopes (about 1% per atomic mass unit) compared to modern air. Such depletion, found previously in well dated samples like 3.5-Gyr-old barite and 3.0-Gyr-old quartz[27,28,32] is proposed to represent the Xe isotope composition of Archaean air.

Building of the model. We used a three-reservoir (mantle crust and atmosphere), first-order rate, box model similar to the one developed by ref. 9 (Supplementary Fig. 4). In such a model, the mantle contained initially primordial noble gases (here, ^{36}Ar) that were subsequently degassed into the atmosphere. However, atmospheric noble gases might have been contributed by sources other than mantle degassing, such as late accretion of volatile-rich bodies. In this case, the model, although conceptually different, yields essentially the same results. Indeed, it does not make a mathematical difference between an early catastrophic event that injects mantle-derived ^{36}Ar into the atmosphere before production of significant ^{40}Ar, and the occurrence of a ^{36}Ar-bearing atmosphere, with later contribution of ^{40}Ar from the mantle. We note that the early degassing event is required by all models based on Ar isotopes, to account for the large ^{40}Ar/^{36}Ar contrast between the mantle and the atmosphere[9]. ^{40}Ar is produced only from the radioactive decay of ^{40}K, with a half-life of 1.25 Gyr. K, initially in the mantle, has been extracted together with Ar, during mantle melting (we assume that both Ar and K are highly incompatible during mantle melting, which is well established in the case of K, and well supported by experimental data for Ar; ref. 31). Ar degasses into the atmosphere and a fraction of K is transferred in the crust (resulting in 20–50% bulk silicate Earth K being stored in the crust now). ^{40}Ar produced in the crust partly degasses (see below). Thus, ^{40}Ar originates from both the mantle and the crust, and its flux into the atmosphere will depend on mantle convection/degassing and also on the volume of crust that stores K. The computations were carried out with the Stella code (http://www.iseesystems.com/softwares/Education/StellaSoftware.aspx). Data used to build this model and constrain its solutions are presented in Supplementary Table 4.

The mantle convection rate directly affects the degassing of Ar and the storage of K in the crust through felsic crust production. To mimic the decrease of heat inside Earth, especially heat due to radioactivity, an exponential decrease of mantle convection is classically assumed[9]. However, such an exponential decrease is not sufficient to explain the modern high ^{40}Ar/^{36}Ar difference between the present-day mantle and atmosphere, so that an early catastrophic degassing event is required that must have occurred in the first 100–200 Myr or so. Such an early high convection rate is independently supported by extinct radionuclides[34–36]. Different durations of catastrophic degassing have been tested and a time interval of 170 Myr gives the largest number of solutions. Thus, degassing rates can be separated into before 170 Myr during the intense degassing with a constant rate, and after 170 Myr with the exponential decrease of degassing. A fraction of radiogenic argon generated in the crust degasses into the atmosphere, with a degassing coefficient taken as variable between 1% and 50%, the last number corresponding to a comparison between Rb–Sr and K–Ar ages for crustal rocks[9]. The model is run with variable atmospheric ^{40}Ar/^{36}Ar ratios constrained by the present quartz data at the different periods of time defined above: $t = 3.5$ Gyr, ^{40}Ar/^{36}Ar = 119–167; $t = 3.0$ Gyr, ^{40}Ar/^{36}Ar = 167–211; and $t = 2.7$ Gyr, ^{40}Ar/^{36}Ar = 190–232. The crustal growth curves obtained with these different closure ages are indistinguishable (Fig. 2), which means that these results are not model-dependent within the 2.7–3.5-Gyr range.

The solutions of our Ar-based model evolution of the continental crust indicate that about half of the present-day continental crust was already present 3.5 Gyr ago (the volume of the Archaean crust is 31%–55% of the present-day volume) and that at our lower age limit of 2.7 Gyr, the crustal volume was 69%–88% of the present-day felsic crust. Our continental crust growth curves (Supplementary Fig. 5) are intermediate between those representing early and intense growth in the Hadean eon[1,3,37] and those representing late[4,38] or sigmoidal growths[5,39]. They predict a larger crustal volume in the Hadean than the model based on the U–Pb or Hf isotope compositions of zircons. However, these geochemical proxies include crustal reworking, which can be corrected for by combining these data with oxygen isotopes[1]. In such a case our model runs are consistent with those derived from the U–Pb, Hf and O isotopes of continental zircons, that is, high crustal production in the Archaean, followed by reduction of the net growth rate by a factor of about 2–4, beginning at around 3.0 Gyr ago (same position of the inflection, Supplementary Fig. 5).

31. Kendrick, M. A., Burgess, R., Pattrick, R. A. D. & Turner, G. Halogen and Ar-Ar age determinations of inclusions within quartz veins from porphyry copper deposits using complementary noble gas extractions. *Chem. Geol.* **177,** 351–370 (2001).
32. Srinivasan, B. Barites: anomalous xenon from spallation and neutron-induced reactions. *Earth Planet. Sci. Lett.* **31,** 129–141 (1976).
33. Meshik, A. P., Hohenberg, C. M., Pravdivtseva, O. V. & Kapusta, Y. S. Weak decay of ^{130}Ba and ^{132}Ba: geochemical measurements. *Phys. Rev. C* **64,** 035205 (2001).
34. Heber, V. S., Brooker, R. A., Kelley, S. P. & Wood, B. J. Crystal-melt partitioning of noble gases (helium, neon, argon, krypton, and xenon) for olivine and clinopyroxene. *Geochim. Cosmochim. Acta* **71,** 1041–1061 (2007).
35. Boyet, M. & Carlson, R. W. ^{142}Nd evidence for early (>4.53 Ga) global differentiation of the silicate. *Earth Sci.* **309,** 576–581 (2005).
36. Caro, G., Bourdon, B., Birk, J. L. & Moorbath, S. ^{146}Sm-^{142}Nd evidence from Isua metamorphosed sediments for early differentiation of Earth's mantle. *Nature* **423,** 428–432 (2003).
37. Fyfe, W. S. Evolution of the Earth's crust: modern plate tectonics to ancient hot spot tectonics? *Chem. Geol.* **23,** 89–114 (1978).
38. Hurley, P. M. Absolute abundance and distribution of Rb, K and Sr in the Earth. *Geochim. Cosmochim. Acta* **32,** 273–283 (1968).
39. Veizer, J. & Jansen, S. L. Basement and sedimentary recycling and continental evolution. *J. Geol.* **87,** 341–370 (1979).

LETTER

doi:10.1038/nature12137

The rewards of restraint in the collective regulation of foraging by harvester ant colonies

Deborah M. Gordon[1]

Collective behaviour, arising from local interactions[1], allows groups to respond to changing conditions. Long-term studies have shown that the traits of individual mammals and birds are associated with their reproductive success[2–6], but little is known about the evolutionary ecology of collective behaviour in natural populations. An ant colony operates without central control, regulating its activity through a network of local interactions[7]. This work shows that variation among harvester ant (*Pogonomyrmex barbatus*) colonies in collective response to changing conditions[8] is related to variation in colony lifetime reproductive success in the production of offspring colonies. Desiccation costs are high for harvester ants foraging in the desert[9,10]. More successful colonies tend to forage less when conditions are dry, and show relatively stable foraging activity when conditions are more humid. Restraint from foraging does not compromise a colony's long-term survival; colonies that fail to forage at all on many days survive as long, over the colony's 20–30-year lifespan, as those that forage more regularly. Sensitivity to conditions in which to reduce foraging activity may be transmissible from parent to offspring colony. These results indicate that natural selection is shaping the collective behaviour that regulates foraging activity, and that the selection pressure, related to climate, may grow stronger if the current drought in their habitat persists.

In ant populations, the colony is the reproductive individual, producing offspring colonies. The study was conducted with a population of about 300 colonies of the red harvester ant, *Pogonomyrmex barbatus*, at a site near Rodeo, New Mexico, USA that has been censused each year since 1985, so the ages of all colonies are known[11]. A colony is founded by a single queen and lives for about 25 years[12–14]. When the colony is about 5 years old, it reaches a stable size of about 10,000 workers[15] and begins to produce reproductives[16], males and gynes, that mate polyandrously. Newly mated gynes found offspring colonies. In a recent study we used microsatellite variation to identify the offspring colonies founded by daughter gynes of parent colonies, and thus to estimate the female component of colony lifetime reproductive success, in the number of offspring colonies founded by daughter gynes[14]. We did not estimate the contribution of males to colony reproductive success. In only about 25% of colonies, daughter gynes successfully founded new colonies, ranging from 1 to 6 offspring colonies per parent colony.

Harvester ant colonies forage for seeds in the desert, where foraging carries a high cost of ant desiccation. Previous work shows that colonies adjust foraging activity to food availability, using interactions between returning and outgoing foragers[17,18], and that colonies vary in the regulation of foraging[12,18]. Foraging activity changes from day to day[17] in response to food supply and humidity, and other conditions such as the number of larvae requiring food. Ants lose water when foraging, and obtain most of their water from metabolizing the fats in the seeds they eat[9,10]. Foraging is regulated using a simple positive feedback system in which outgoing foragers are stimulated to leave the nest when they interact with returning foragers carrying food into the nest[19]. The rate of forager return reflects current food supply because each forager searches until it finds a seed[20], so foragers return more quickly the more food is available. Harvester ant colonies vary in the regulation of foraging activity, by varying in the response to the rate of forager return[17,18]. Colonies show characteristic foraging behaviour from year to year[12], reflecting colony-specific behavioural reaction norms[8] for the relation between foraging activity and current conditions.

How a colony regulates its foraging behaviour is associated with its lifetime reproductive success. In poor conditions when humidity is low, foraging activity reflects reproductive success more strongly than when humidity is high. Foraging activity in colonies with and without offspring colonies differed overall on dry days (Cochran-Mantel-Haenszel test, $M^2 = 10.96$, d.f. $= 2$, $P = 0.004$) but not on humid days (Cochran-Mantel-Haenszel test, $M^2 = 0.27$, d.f. $= 2$, $P = 0.87$) (Fig. 1). Of the colonies that foraged at all on dry days, more colonies with than without offspring colonies tended to show low rather than high foraging activity (Fig. 1), although the difference between colonies with and without offspring colonies was not significant (Mantel-Haenszel chi-squared test, NS).

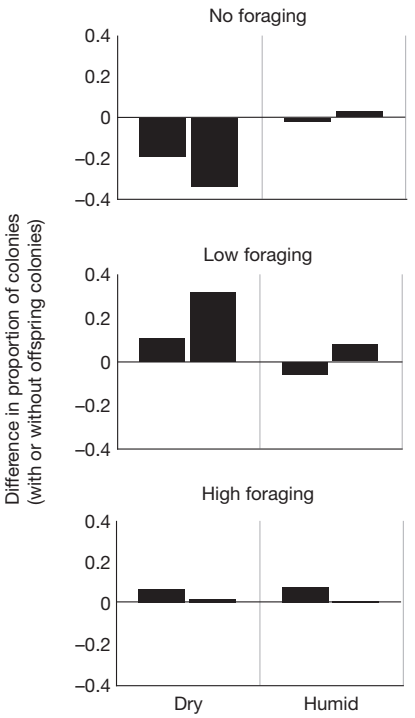


Figure 1 | Foraging activity on dry days is associated with reproductive success. Each bar shows foraging activity on one day, for two dry and two humid days in 2012 (weather data are in Supplementary Table 1). Each bar shows the difference obtained by subtracting the proportions of colonies with and without offspring colonies, showing foraging activity in the indicated category: none, low, or high. The difference shown is the proportion of 37 colonies with offspring colonies, minus the proportion of 24 colonies without offspring colonies.

[1]Department of Biology, Stanford University, Stanford, California 94305-5020, USA.

Stability in foraging activity, in good conditions, is associated with high reproductive success. Colonies with offspring colonies fluctuated less in foraging activity over 5 humid days in 2011 than those without offspring colonies, showing a higher ratio of smallest to highest foraging rate (colonies with offspring colonies, mean ratio = 0.51 (s.d. = 0.17); colonies without offspring colonies, mean ratio = 0.36 (s.d. = 0.17); $t = 2.09$; $P < 0.049$, two-tailed t-test). The standard deviation of a colony's foraging rate, in numbers of returning foragers per 30 s, over 5 humid days, was lower for colonies with offspring colonies ($n = 21$) than for colonies without offspring colonies ($n = 21$) ($t = 2.60$, $P < 0.01$, two-tailed t-test). The mean and sum of foraging rate over 5 days was not significantly different (colonies with offspring colonies, mean = 32.7 (s.d. = 16.5) ants returning per 30 s, colonies without offspring colonies, mean = 34.5 (s.d. = 22.7) ants returning per 30 s, $t = 0.29$, NS; colonies with offspring colonies, mean sum foraging = 163.7 (s.d. = 82.2) ants returning per 30 s, colonies without offspring colonies, mean sum foraging = 172.6 (s.d = 113.5) ants returning per 30 s, $t = 0.3$, NS).

There was no survival cost of not foraging (Fig. 2). Previous work has demonstrated that a colony does not forage every day, and the proportion of days that it forages actively is a colony-specific trait that persists from year to year[12]. The proportion of days a colony foraged ranged from 0.32 to 1.0 in 1986 and from 0.35 to 1.0 in 1987, and colony age at death ranged from 7 to 30 years. There was no relation between the proportion of days that a colony foraged in 1986 or 1987 and its age at its death sometime in the subsequent 25 years (1986, $z = -1.227$, d.f. = 32, $P = 0.21$; 1987, $z = 0.77$, d.f. = 35, $P = 0.4$, Spearman's rank correlation). It appears that colonies can collect sufficient food on good days to desist from foraging on poor days without risking starvation. Although colonies compete with neighbours for foraging area[21], and food is apparently a limiting resource for desert granivores[22], colonies can store seeds for long periods, up to many months[23].

There is some evidence that the regulation of foraging may be transmissible from parent queens to their daughter queens. Because daughter queens do not tend to found colonies near their parents[14], there is no contact between parent and offspring colonies that could lead to cultural transmission of collective behaviour. The 42 offspring colonies of 17 parent colonies resembled their parents in the choice of days in which to reduce foraging activity. In the course of 5 days in 2011, 11 of 17 parent colonies reduced foraging activity on the same day, and the offspring colonies of 5 of these did so as well. All 6 of the parent colonies that chose an uncommon day to reduce foraging had offspring colonies that also chose an uncommon day. This produced a significant association between parents and offspring colonies in the choice of day on which foraging was most reduced (Fisher's exact test, $P = 0.04$). This indicates that offspring colonies may resemble their parent colonies in the reaction norm that links particular conditions, characteristic of a certain day, to the reduction of foraging activity. However, there was no correlation between parents and offspring in the sum ($z = -0.83$, d.f. = 15, $P = 0.4$) or standard deviation ($z = -0.63$, d.f. = 15, $P = 0.5$) of foraging over 5 days. Many factors probably produce variation among colonies in foraging activity, such as variation in the amount of stored food and in the number of brood to feed[24]. If there is heritable variation among colonies in sensitivity to day-to-day changes in weather conditions, data from many colonies on many days that differ greatly in weather conditions might be needed to discern a correlation in the foraging behaviour of parent and offspring colonies.

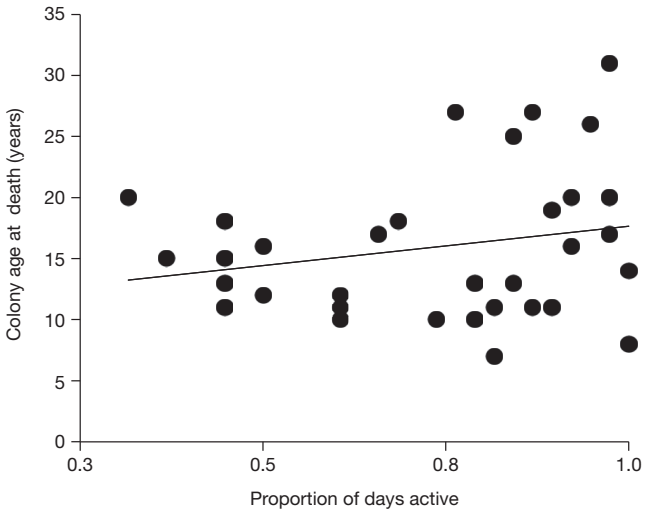


Figure 2 | No survival cost of not foraging. Each point shows the proportion of days out of 38 days in 1986 that a colony foraged actively and the age of the colony at death. The line shows the least-squares fit; colony age at death and proportion of days active were not significantly correlated.

That some aspect of foraging behaviour may be transmissible from queens to daughter queens is consistent with previous work indicating that foraging behaviour is transmissible from queens to daughter workers. A queen of *P. barbatus* can live for about 25 years, whereas workers, her daughters, live at most about a year[25]. Thus colony-specific foraging behaviour that persists from year to year is due to characteristics that appear in successive years in distinct, successive cohorts of workers, all of which are daughters of the same queen, though not of the same fathers. Variation in the foraging and circadian genes whose expression is associated with foraging activity in this species[26] may lead to the transmissibility of foraging activity.

It may seem surprising that high reproductive success is not associated with high foraging activity. In much of foraging theory, the amount of food collected is assumed to be correlated with reproductive success. In studies of social insects, the assumption that more food means more offspring arises from a chain of inference: the more workers, the more food is collected; the more food, the more reproductives can be produced; and the higher the reproductive output, the greater the realized reproductive success of the colony in offspring colonies. Both the measures of the variables and the chain of inference itself require testing. Here, because we can estimate colony lifetime reproductive success, it was possible to test directly whether in fact reproductive success, in offspring colonies founded by daughter queens, is correlated with foraging activity. Colonies that forage actively on more days do not live longer; colonies with or without offspring colonies differ most in poor conditions, when colonies with high reproductive success tend to show low foraging activity. Although it is clear that a colony with inadequate food could not survive or make reproductives at all, it seems that once some minimum threshold of food supply is reached, other factors, including perhaps the cost of desiccation, have a stronger impact on colony reproductive success than persistently high foraging activity. Like many animal species that store food, for example in fat reserves, harvester ant colonies store seeds for many months[23]. Harvester ant colonies that conserve more water may be able to produce more, or better-hydrated female reproductives that can survive longer during the founding stage[27]. In other conditions, such as in tropical forests where the cost of foraging is low, other constraints, such as interspecific competition, probably create different evolutionary pressures on the collective foraging behaviour of ant colonies.

METHODS SUMMARY

Foraging behaviour was observed in a population of about 300 colonies of *P. barbatus* at a site near Rodeo, New Mexico, USA in which the ages of all colonies have been determined, and the female component of lifetime reproductive success has been estimated for most mature colonies[14]. To test the relation of foraging activity and survival, correlations were examined between the foraging activity of 34 colonies on 38 days in 1986 and 37 colonies on 34 days in 1987, and the number of years the colony survived. To compare foraging activity of colonies with and without offspring colonies, foraging activity was compared for 21 colonies that had

offspring colonies and 21 that did not on 5 humid days in 2011, and for 37 colonies with and 24 colonies without offspring colonies on 2 dry and 2 humid days in 2012. The transmissibility of foraging behaviour from parent to offspring colonies was evaluated by examining the association between 17 parent and 42 offspring colonies in sensitivity to conditions in which to reduce foraging, and for the same 17 parent and 42 offspring colonies, the correlation over 5 days in 2011 of mean parent and offspring colony values for foraging rate and standard deviation of foraging rate.

Full Methods and any associated references are available in the online version of the paper.

Received 1 February; accepted 2 April 2013.
Published online 15 May 2013.

1. Pratt, S. C. & Sumpter, D. J. T. A tunable algorithm for collective decision-making. *Proc. Natl Acad. Sci. USA* **103,** 15906–15910 (2006).
2. Clutton-Brock, T. & Sheldon, B. C. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends Ecol. Evol.* **25,** 562–573 (2010).
3. Garant, D. Kruuk, L. E. B., McCleery, R. H. & Sheldon, B. C. Evolution in a changing environment: a case study with great tit fledging mass. *Am. Nat.* **164,** E115–E129 (2004).
4. Altmann, S. A. Diets of yearling female primates (*Papio cynocephalus*) predict lifetime fitness. *Proc. Natl Acad. Sci. USA* **88,** 420–423 (1991).
5. Nussey, D. H., Wilson, A. J. & Brommer, J. E. The evolutionary ecology of individual phenotypic plasticity in wild populations. *J. Evol. Biol.* **20,** 831–844 (2007).
6. Moyes, K. *et al.* Exploring individual quality in a wild population of red deer. *J. Anim. Ecol.* **78,** 406–413 (2009).
7. Gordon, D. M. *Ant Encounters: Interaction Networks and Colony Behavior*. (Princeton Univ. Press, 2010).
8. Dingemanse, N. J., Kazem, A. J., Réale, D. & Wright, J. Behavioural reaction norms: animal personality meets individual plasticity. *Trends Ecol. Evol.* **25,** 81–89 (2010).
9. Lighton, J. R. & Bartholomew, G. A. Standard energy metabolism of a desert harvester ant, *Pogonomyrmex rugosus*: effects of temperature, body mass, group size, and humidity. *Proc. Natl Acad. Sci. USA* **85,** 4765–4769 (1988).
10. Lighton, J. R. B. & Feener, D. H., Jr. Water-loss rate and cuticular permeability in foragers of the desert ant *Pogonomyrmex rugosus. Physiol. Zool.* **62,** 1232–1256 (1989).
11. Gordon, D. M. & Kulig, A. W. Founding, foraging and fighting: colony size and the spatial distribution of harvester ant nests. *Ecology* **77,** 2393–2409 (1996).
12. Gordon, D. M. Behavioral flexibility and the foraging ecology of seed-eating ants. *Am. Nat.* **138,** 379–411 (1991).
13. Gordon, D. M. & Kulig, A. W. The effect of neighboring colonies on mortality in harvester ants. *J. Anim. Ecol.* **67,** 141–148 (1998).
14. Ingram, K. K., Pilko, A., Heer, J. & Gordon, D. M. Colony life history and lifetime reproductive success of red harvester ant colonies. *J. Anim. Ecol.* **82,** 540–550 (2013).
15. Gordon, D. M. How colony growth affects forager intrusion in neighboring harvester ant colonies. *Behav. Ecol. Sociobiol.* **31,** 417–427 (1992).
16. Gordon, D. M. The development of an ant colony's foraging range. *Anim. Behav.* **49,** 649–659 (1995).
17. Gordon, D. M., Holmes, S. & Nacu, S. The short-term regulation of foraging in harvester ants. *Behav. Ecol.* **19,** 217–222 (2008).
18. Gordon, D. M., Guetz, A., Greene, M. J. & Holmes, S. Colony variation in the collective regulation of foraging by harvester ants. *Behav. Ecol.* **22,** 429–435 (2011).
19. Prabhakar, B., Dektar, K. N. & Gordon, D. M. The regulation of ant colony foraging activity without spatial information. *PLoS Comp. Biol.* **8,** e1002670 (2012).
20. Beverly, B. D., McLendon, H., Nacu, S., Holmes, S. & Gordon, D. M. How site fidelity leads to individual differences in the foraging activity of harvester ants. *Behav. Ecol.* **20,** 633–638 (2009).
21. Adler, F. R. & Gordon, D. M. Optimization, conflict, and non-overlapping foraging ranges in ants. *Am. Nat.* **162,** 529–543 (2003).
22. Davidson, D. W. Some consequences of diffuse competition in a desert ant community. *Am. Nat.* **116,** 92–105 (1980).
23. Gordon, D. M. The spatial scale of seed collection by harvester ants. *Oecologia* **95,** 479–487 (1993).
24. Cassill, D. L. & Tschinkel, W. R. Allocation of liquid food to larvae via trophallaxis in colonies of the fire ant, *Solenopsis invicta. Anim. Behav.* **50,** 801–813 (1995).
25. Gordon, D. M. & Holldobler, B. Worker longevity in harvester ants. *Psyche (Stuttg.)* **94,** 341–346 (1987).
26. Ingram, K. K., Kleeman, L. & Peteru, S. Differential regulation of the foraging gene associated with task behaviors in harvester ants. *BMC Ecol.* **11,** 19 (2011).
27. Johnson, R. A. & Gibbs, A. G. Effect of mating stage on water balance, cuticular hydrocarbons and metabolism in the desert harvester ant, *Pogonomyrmex barbatus*. *J. Insect Physiol.* **50,** 943–953 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements Thanks to the many people who helped with field work: in 1986 and 1987, K. Roth; in 2011, X. Ampuero, K. Dektar, M. Greene, J. Hickman, A. Merrell and N. Pinter-Wollman; in 2012, J. Queirolo, J. Rasiel, C. Wayne; in both 2011 and 2012, S. Crow, L. Howard and E. Pless. Many thanks to M. Coram for statistical advice and help. I am grateful to D. Kennedy and J. Ober for helpful discussions and to M. Feldman and W. Flesch for comments on the manuscript. The work was funded by the Stanford Office of the Dean of Research, Stanford Emergence of Cooperation Project and the National Science Foundation grant IOS-0718631.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.M.G. (dmgordon@stanford.edu).

METHODS

Number of days foraging and colony survival. A harvester ant colony does not forage actively every day, and the extent of a colony's tendency to forage actively persists from year to year[12]. The survival cost of not foraging was examined using the correlation between number of days active and colony survival over the subsequent 25 years. We recorded for 34 colonies in 1986 whether the colony was foraging actively on each of 38 days (between 24 Jun to 6 Aug 1986), and for 37 colonies in 1987 whether it was foraging actively on each of 34 days (between 5 July to 19 Aug 1987). Foraging activity was recorded in 23 of these colonies in both years. All but 3 of the colonies had died by the 2011 census. The number of years the colony lived was the number of years from the year the colony was founded until the year it was determined to be dead. The oldest colonies were determined to be at least 5 when the census began in 1985, but may have been older. The relation between foraging activity and colony survival was examined by using Spearman's rank correlation to test for a correlation between the proportion of days (out of 38 in 1986 and out of 34 in 1987) that a colony foraged actively, and the number of years it lived.

General methods for measures of foraging activity. Supplementary Table 1 shows temperature, dew point and relative humidity for the days on which foraging was measured.

For measurement of foraging activity, Supplementary Table 2 lists which colonies were observed in each year and the colony age in 2010. There is no evidence of reproductive senescence in this species[14], and age-specific fecundity is approximately the same from age 5 years onward[14]. Only colonies aged 10 years or older in 2010, when parent–offspring colony pairs were identified[14], were used in comparisons of colonies with and without offspring colonies, so all colonies had at least 5 years (from ages 5 to 10 years) in which to produce offspring colonies.

Foraging rate in humid conditions. In 2011, foraging rate was measured as the number of ants returning to the nest in 30 s, on 5 humid days in August (Supplementary Table 1) in 42 colonies aged 12 years or older, of which 21 had offspring colonies and 21 had no offspring colonies. To test whether foraging rate is associated with reproductive success, I determined for each colony: (1) the sum of the foraging rates over the 5 days; (2) the mean foraging rate over the 5 days; (3) the standard deviation in foraging rate over the 5 days; and (4) a measure of the extent to which the colony ever decreased its foraging rate. Of the 42 colonies observed, 12 colonies with offspring colonies and 11 colonies without offspring colonies foraged actively on all 5 days. For each of these 23 colonies, the smallest normalized foraging rate was calculated, representing the lowest proportion observed to forage in the course of the 5 days. Foraging rate was normalized for differences among colonies in colony size by dividing each day's foraging rate by the largest rate observed in that colony in the course of the 5 days. For these 4 measures, two-tailed *t*-tests were used to compare colonies with and without offspring colonies.

2012 comparison of foraging rate in dry and wet days. In August 2012, during a time of severe drought, foraging activity was measured on two dry days (12 August and 13 August) and two humid days (23 August and 24 August) in 24 colonies that had no offspring colonies, and 37 colonies with offspring colonies, with ages 10 to 30 years. Foraging activity was ranked, based on the range of foraging rates previously observed[18,19], as none, low (1 to 4 returning foragers per 5 s), or high (5 or more returning foragers per 5 s). Cochran-Mantel-Haenszel tests were used to test for a difference between the two dry and the two humid days in the proportions of colonies with and without offspring colonies showing each foraging rate (none, low or high). Then Mantel-Haenszel chi-squared tests were used to test separately for a difference in the proportions of colonies with and without offspring colonies showing none, low or high foraging, stratified by dry or wet days.

Transmissibility of colony foraging behaviour. In Aug 2011, foraging rate was measured as described in comparison of foraging rate in dry and wet days for the same 5 days in an additional 19 colonies for a total of 61 colonies. Of these colonies, 17 were parent colonies, ranging in age from 10 to 30 years old, and 44 colonies, ranging in age from 3 to 29 years old, were one of 1 to 5 offspring colonies founded by a daughter queen of one of the parent colonies. To evaluate the transmissibility of foraging behaviour from parent to offspring colony, Spearman's rank correlation tests were used to examine the correlation, between the value for the parent colony and the mean value for all of that parent colony's offspring colonies, of the sum over the 5 days of foraging rate and of the standard deviation over the 5 days of foraging rate.

I also examined the similarity of parent and offspring colonies in the choice of day in which it most reduced foraging activity. Although all 5 days were fairly humid (Supplementary Table 1), conditions and foraging activity all differed from day to day. I found for each colony that was active on all 5 days (17 parents and 42 offspring colonies), the day or days on which the foraging rate was lowest; any day with a foraging rate within 1 ant per second of the lowest day's foraging rate was also considered a day on which foraging rate was lowest. Foraging activity was lowest on August 11 for 65% of parent colonies and 31% of offspring colonies. Each parent colony was classified as having its lowest foraging rate either on August 11 or on some other days (including another day as well as August 11). For each parent colony, the lowest foraging rate of half or more of its offspring colonies was determined to be August 11 or on some other day. A Fisher's exact test was used to determine whether parent colonies that had lowest foraging rates on a day other than the most common day were likely to have offspring colonies that also had lowest foraging rates on a day other than the most common day.

LETTER

OPEN

doi:10.1038/nature12132

Architecture and evolution of a minute plant genome

Enrique Ibarra-Laclette[1], Eric Lyons[2], Gustavo Hernández-Guzmán[1,3], Claudia Anahí Pérez-Torres[1], Lorenzo Carretero-Paulet[4], Tien-Hao Chang[4], Tianying Lan[4,5], Andreanna J. Welch[4], María Jazmín Abraham Juárez[6], June Simpson[6], Araceli Fernández-Cortés[1], Mario Arteaga-Vázquez[7], Elsa Góngora-Castillo[8], Gustavo Acevedo-Hernández[9], Stephan C. Schuster[10,11], Heinz Himmelbauer[12,13], André E. Minoche[12,13,14], Sen Xu[15], Michael Lynch[15], Araceli Oropeza-Aburto[1], Sergio Alan Cervantes-Pérez[1], María de Jesús Ortega-Estrada[1], Jacob Israel Cervantes-Luevano[1], Todd P. Michael[16], Todd Mockler[17], Douglas Bryant[17], Alfredo Herrera-Estrella[1], Victor A. Albert[4] & Luis Herrera-Estrella[1]

It has been argued that the evolution of plant genome size is principally unidirectional and increasing owing to the varied action of whole-genome duplications (WGDs) and mobile element proliferation[1]. However, extreme genome size reductions have been reported in the angiosperm family tree. Here we report the sequence of the 82-megabase genome of the carnivorous bladderwort plant *Utricularia gibba*. Despite its tiny size, the *U. gibba* genome accommodates a typical number of genes for a plant, with the main difference from other plant genomes arising from a drastic reduction in non-genic DNA. Unexpectedly, we identified at least three rounds of WGD in *U. gibba* since common ancestry with tomato (*Solanum*) and grape (*Vitis*). The compressed architecture of the *U. gibba* genome indicates that a small fraction of intergenic DNA, with few or no active retrotransposons, is sufficient to regulate and integrate all the processes required for the development and reproduction of a complex organism.

Like other carnivorous plants, *Utricularia* (Lentibulariaceae) species derive nitrogen and phosphorus supplements by trapping and digesting prey organisms[2,3]. Lentibulariaceae are asterid angiosperms closely related to the model plants snapdragon (*Antirrhinum*) and monkey flower (*Mimulus*). Among *Utricularia* species, the intricate, water-filled suction bladders are variously arrayed on plant parts, and may even take the place of an embryonic leaf[2,4]. Whereas *Utricularia* vegetative structures are extremely diverse, its snapdragon-like flowers are stereotypical for plants of its asterid clade[2] (Fig. 1a). Interestingly, these inhabitants of nutrient-poor environments do not bear true roots[4].

Our *U. gibba* genome assembly, produced using a hybrid (454/Illumina/Sanger) sequencing strategy, closely matches the genome size estimated by flow cytometry (77 megabases (Mb)) (Supplementary Information section 1). Remarkably, despite its tiny size, the (G+C)-rich *U. gibba* genome accommodates about 28,500 genes, slightly more than *Arabidopsis*, papaya, grape or *Mimulus*, but less than tomato (Supplementary Information section 2). Indeed, the *U. gibba* genome has experienced a small, approximately 1.5% net gain across a conserved set of single-copy genes[5] (Supplementary Information section 2.6). Synteny analysis reveals that *U. gibba* has undergone three sequential WGD events since last common ancestry with tomato and grape, with one of these duplications possibly shared by the closely related species *Mimulus* (Fig. 1a and Supplementary Information section 7). Consequently, the *U. gibba* genome seems to be 8× with respect to the palaeohexaploid (3×) core eudicot ancestor[6] (Fig. 1b), whereas *Arabidopsis* is 4× with a genome 1.5-times larger[7]. Compared with independently polyploid tomato[8], the *U. gibba* genome shows extremely fractionated gene loss (Fig. 1c), with almost two-thirds of syntenic genes shared with tomato having returned to single copy (Supplementary Information section 7.4 and Supplementary Table 39).

Intergenic sequence contraction in the *U. gibba* genome is particularly apparent in the paucity of repetitive DNA and mobile elements (Supplementary Table 8). Whereas repetitive DNA accounts for 10–60% of most plant genomes, in *U. gibba* it only amounts to 3%, including 569 mobile elements (Supplementary Information section 2). Notably, retrotransposable elements, which largely dominate angiosperm genomes, are rare in the *U. gibba* genome; we identified only 379, amounting to about 2.5% of the genome. Of these, only 95 seem complete and therefore potentially capable of further retrotransposition (Supplementary Information section 2.1 and Supplementary Tables 8 and 9). We found that all genes known to be involved in retrotransposon silencing have homologues in *U. gibba* (Supplementary Table 28), as well as a set of 75 microRNAs (miRNAs) belonging to 19 families (Supplementary Table 29 and Supplementary data 7). These results indicate that, despite its small genome, the general repertoire of miRNA-mediated gene regulation mechanisms in plants is conserved in *U. gibba* (Supplementary Table 29). Together, these data indicate that any influence of retrotransposon proliferation on *U. gibba* genome size must be countered by fractionation after WGDs and also by the silencing of these mobile elements.

The *U. gibba* genome contains a high percentage of small, putative promoters (Supplementary Fig. 11 and Supplementary Data 5) and tail-to-tail gene pairs with overlapping 3′ ends (Supplementary Tables 25 and 26). This configuration is similar to, but about 50% shorter than, that in *Arabidopsis*, which has led to denser packing in *U. gibba* gene islands (Fig. 2a). Using transient expression analysis, we confirmed that several short intergenic sequences function as transcriptional promoters, including a 400-base-pair region serving as a bidirectional promoter of a head-to-head gene pair (Fig. 2b and Supplementary Information section 3). These results indicate that the binding sites for transcription factors that direct the expression of *U. gibba* genes remain in their 5′ flanking regions, and that conserved *cis*-acting elements are compressed in at least a portion of the promoters of this carnivorous plant (Supplementary Fig. 11). Genome size contraction is also reflected at the level of introns, which showed smaller size and a slightly reduced number per gene (Supplementary Information section 5).

Compressed promoter spaces, fewer exons per gene than *Arabidopsis* (that is, net intron loss; Supplementary Table 12), and missing segments

[1]Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), 36821 Irapuato, Guanajuato, México. [2]The School of Plant Sciences and iPlant Collaborative, University of Arizona, Tucson, Arizona 85721, USA. [3]Departamento de Alimentos, División de Ciencias de la Vida, Universidad de Guanajuato, 36500 Irapuato, Guanajuato, México. [4]Department of Biological Sciences, University at Buffalo, Buffalo, New York 14260, USA. [5]Department of Biology, Chongqing University of Science and Technology, 4000042 Chongqing, China. [6]Departamento de Genética, Unidad Irapuato, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional (CINVESTAV), 36821 Irapuato, Guanajuato, México. [7]Instituto de Biotecnología y Ecología Aplicada, Universidad Veracruzana, 91090 Xalapa, Veracruz, México. [8]Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, USA. [9]Centro Universitario de la Ciénega, Universidad de Guadalajara, 47840 Ocotlán, Jalisco, México. [10]Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [11]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, 637551 Singapore. [12]Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain. [13]Universitat Pompeu Fabra (UPF), 08018 Barcelona, Spain. [14]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany. [15]Department of Biology, Indiana University, Bloomington, Indiana 47405, USA. [16]Waksman Institute of Microbiology and Department of Plant Biology and Pathology, Rutgers University, New Brunswick, New Jersey 08854, USA. [17]The Donald Danforth Plant Science Center, St. Louis, Missouri 63132, USA.

or whole genes in retroelements (Supplementary Fig. 4) support the notion that numerous microdeletions have occurred during *U. gibba* genome evolution, as previously observed in *Arabidopsis*[9] and maize[10]. Furthermore, the presence of numerous solo long terminal repeat (LTR) elements (a single copy of an LTR that is the product of homologous recombination events between two identical or related LTR-retrotransposons) in the *U. gibba* genome (Fig. 2c and Supplementary Fig. 5) indicates that large-scale recombinational deletions have also occurred[11]. Unlike the contracted nuclear genome, the plastid and mitochondrial genomes of *U. gibba* are quite similar in structure to those of other angiosperms (Supplementary Information section 8 and Supplementary Figs 35–38) with no apparent shortening of intergenic regions (Supplementary Tables 41 and 43). Therefore, the evolutionary forces acting to reduce *U. gibba* genome size seem to have affected only the nucleus.

We investigated the coding DNA content of the *U. gibba* genome compared to the *Arabidopsis*, tomato, grape, *Mimulus* and papaya genomes in two complementary ways: (1) by predicted protein domains, and (2) by gene family classification. In the first approach, we compared protein domains and applied a likelihood ratio test to examine the significance of difference in numbers of Pfam domains (Supplementary Table 15). 97% of domain groups did not show significant differences among the plant species analysed, and of the remaining 3%, only 40% represented instances where *U. gibba* had fewer domain members than other plant species (Supplementary Table 16).

To gain insight into specific differences in the genic repertoire of *U. gibba* and their potential biological significance, in the second approach we classified gene families in the *U. gibba*, *Arabidopsis*, tomato, grape and papaya genomes using OrthoMCL[12]. Out of a total of 18,991 gene families, 1,275 have no *U. gibba* members (57% representing single-gene families, Supplementary Table 18), whereas 1,804 showed an increased number of genes in *U. gibba* (Supplementary Table 19). Several gene families specifically lost or conspicuously reduced in *U. gibba* may have functions related to its unusual embryogenesis (frequently involving asymmetrical production of shoot apical organs and absence of true cotyledons), its frequent shoot–leaf indistinction, and its lack of true roots (Supplementary Table 18; see references in Supplementary Information section 2.5). These include homologues of *AT1G68170* (a nodulin MtN21-like transporter, differentially expressed in globular-stage embryos and cotyledons), *PEI1* (an embryo-specific zinc finger transcription factor required for heart-stage embryo formation), and a paralogue of *FD* (involved in flowering but also expressed in embryos and cotyledons). In addition, compared to the two to three member gene family in all other species examined, *U. gibba* contains a single member of the *CASPARIAN STRIP MEMBRANE DOMAIN PROTEIN* family, which encodes proteins involved in Casparian strip formation in *Arabidopsis* roots. Other genes missing in *U. gibba* may also be involved in root development and physiology: homologues of *WAK* (a cell-wall-associated Ser/Thr kinase involved in cell elongation and lateral root development), *NAXT1* (a nitrate efflux transporter mainly expressed in the cortex of adult roots), *MYB48* and *MYB59*

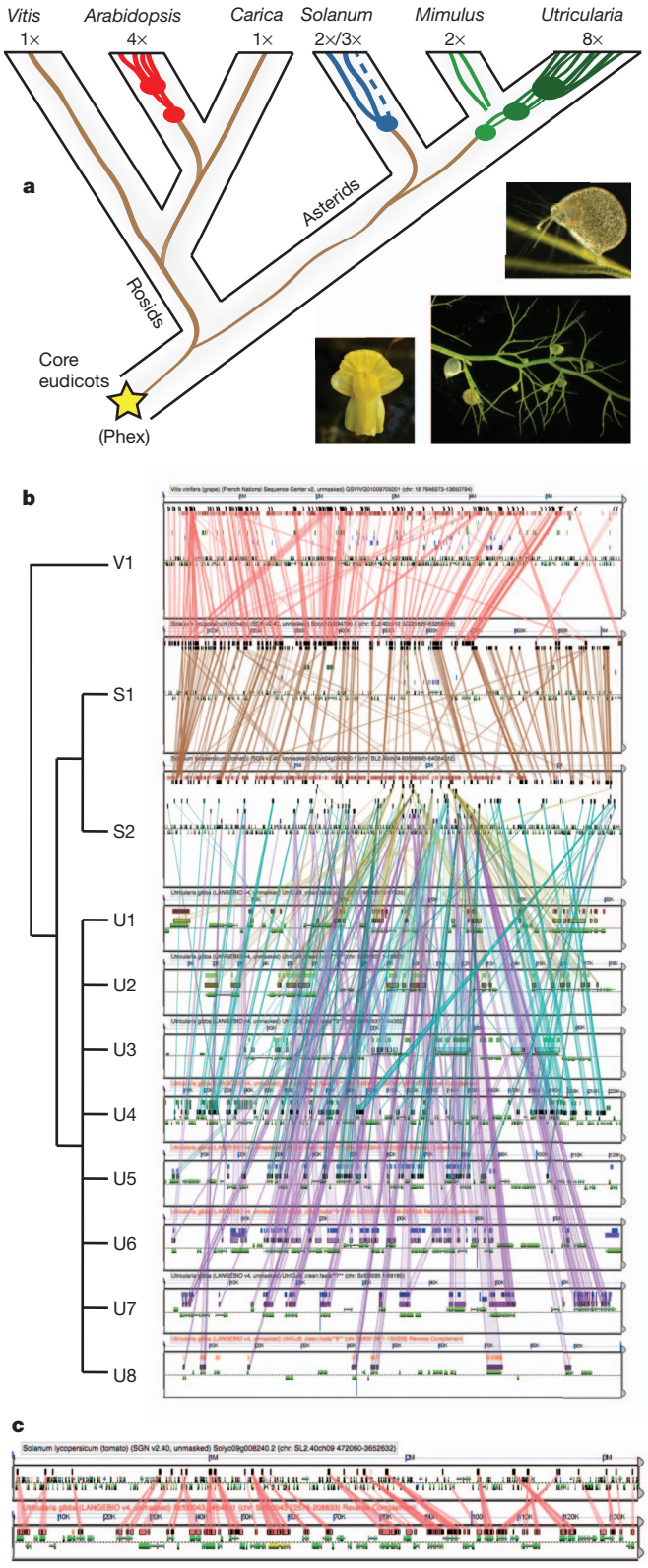


Figure 1 | Syntenic analysis of the *Utricularia gibba* genome. **a**, Whole-genome duplication (WGD) history highlighting the phylogenetic position of *U. gibba*. *Vitis*, *Arabidopsis* and *Carica papaya* are rosids; *Arabidopsis* has had two WGDs since the paleohexaploid (Phex) core eudicot ancestor. Tomato (*Solanum*), *Mimulus* and *U. gibba* are asterids; tomato has a mix of duplicated and triplicated regions; *U. gibba* has had three WGDs since common ancestry with tomato and the Phex ancestor. *Mimulus* has had a single WGD[25] that may also be the most ancient WGD observed for *U. gibba* (see Supplementary Information section 7.1.3). *U. gibba* flowers are similar to those of *Mimulus* (that is, like snapdragons); tiny suction traps are borne on highly divided branching structures (insets, clockwise from left). **b**, A microsyntenic analysis shows that *U. gibba* (U) is 8:2:1 relative to homologous tomato (T) and *Vitis* (V) regions, respectively. As such, *U. gibba* is a 16-ploid with respect to *Vitis*, and the polyploidy of tomato is entirely independent (Supplementary Information section 7). Coloured lines connect high-scoring segment pairs (HSPs) on genome blocks masked for non-coding sequences. Gene models lie in the centres of each block, below the HSPs. This analysis may be regenerated by CoGe at http://genomeevolution.org/r/4wvh. **c**, Fractionation in a given *U. gibba* region can be massive with respect to tomato; the regions shown include an over 3 Mb block of the tomato genome (top), strongly syntenic and colinear to an approximately 130-kb block of *U. gibba*, representing an approximately 20:1 difference in total DNA. This analysis may be regenerated by CoGe at http://genomevolution.org/r/5cet.

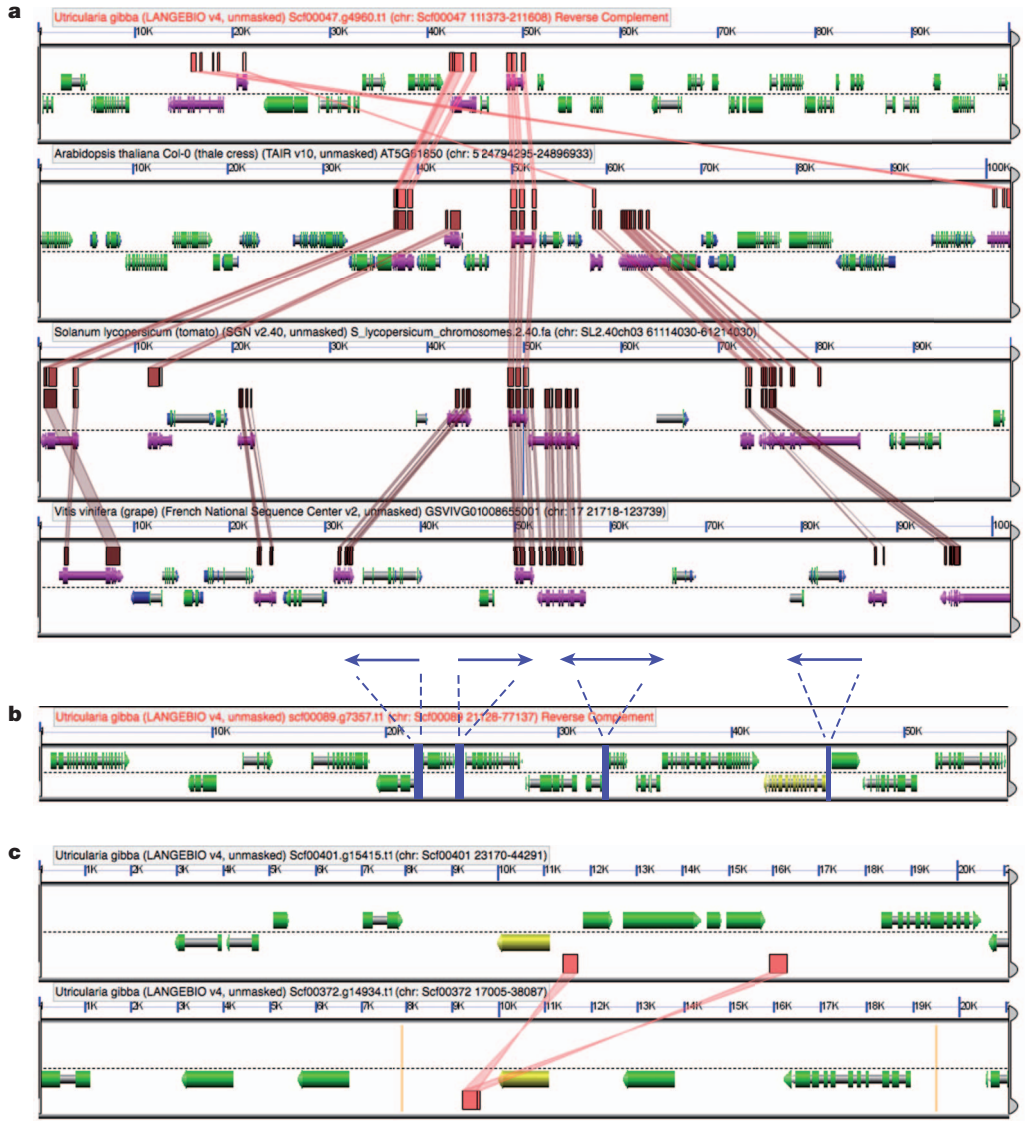


Figure 2 | Architecture of the *Utricularia gibba* genome. **a**, *U. gibba* gene islands are more compact than in *Arabidopsis*, and much higher in gene density than tomato or grape. For example, the *Arabidopsis LEAFY* gene lies directly in the middle of the second block from the top, which is an approximately 100-kb region from *Arabidopsis* chromosome 5. There are 28 genes in this view. In the corresponding *U. gibba* block (top), there are 34 genes within the same-sized region, which is therefore approximately 18% more densely packed. In tomato (3rd block) and grape (4th), there are many fewer genes (14 and 17, respectively) for a much lower density of gene space. **b**, Promoter spaces in *U. gibba* can be very short. Shown is part of a scaffold (scf00089), the sequence of which was verified by PCR walking. Four promoter regions (blue) showed reproducible activity in transient expression experiments (see Supplementary Information section 3). For example, the short bidirectional promoter between a divergent gene pair is approximately 400 bp. Other gene arrangements, tandem and convergent, can be seen in this example. **c**, Solo LTR remains of ectopically recombined mobile elements can be identified in the *U. gibba* genome. This example shows two blocks from *U. gibba*, the Solo LTR in the bottom block being homologous to the LTR pair present in the top block. In **a**, syntenic HSPs are shown as coloured lines connecting particular gene models (purple). Results from **a** and **c** can be regenerated at http://genomevolution.org/r/5kv5 and http://genomevolution.org/r/8lvv, respectively. See Supplementary Information for further discussion of **b** and **c**.

(nitrogen-responsive genes involved in the regulation of cell cycle progression and root growth), and the MADS box genes *ANR1* (*ARABIDOPSIS NITRATE REGULATED 1* (*AGL44*)) and *XAL1* (*XAANTAL1* (*AGL12*)). *ANR1* is a component of a signalling pathway that regulates lateral root growth in response to external NO_3 supply, whereas *XAL1* is involved in root-cell differentiation and flowering time. At least 50 MADS box genes are known to be expressed in *Arabidopsis* roots, of which the *AGL17*-like type II clade is noteworthy as all its members are expressed in roots, and four of them (*AGL16*, *AGL17*, *AGL21* and *AGL44*) have been reported to be root-specific, as are the type I genes *AGL26* and *AGL56*. Interestingly, contractions and losses in all of these root-expressed MADS box gene clades/subfamilies account for much of the global reduction of the MADS box gene family in *U. gibba* (Supplementary Fig. 7). In contrast, other MADS box gene subfamilies were found to be specifically expanded in *U. gibba* (see references in Supplementary Information section 2.5). One such example is *SOC1*, a gene expressed in shoots with a well-characterized role in regulating flowering time and a possible role in response to phosphorus and sulphur (but not nitrogen) availability. Because it has been reported in *Utricularia vulgaris* that trap formation is induced by low phosphorus availability but not by low nitrogen[13], it is possible that the marked expansion of the *U. gibba* SOC1-like clade is related to the adaptive capacity for phosphorus scavenging from trapped prey. Three clusters representing members of different TCP (TEOSINTE BRANCHED1/CYCLOIDEA/PCF) transcription factor clades are also expanded in *U. gibba*. These genes regulate plant morphogenesis, including branching, and it is tempting to speculate that specific clade expansions may be related to the genus-wide diversity of branching patterns in *Utricularia*[2].

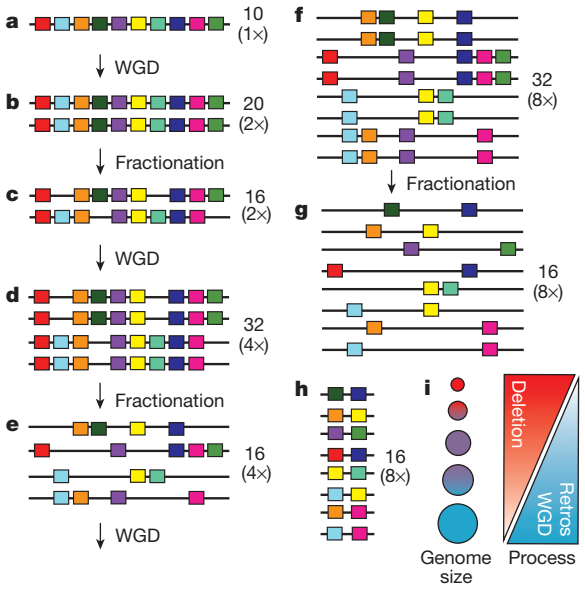


Figure 3 | A model of genome size reduction and the plant genome size evolutionary spectrum. **a**, The initial diploid genome has 10 genes. **b**, **c**, After one WGD (**b**), there are 20 genes in the tetraploid, which fractionate into 16 genes (**c**). **d**–**g**, After another round of WGD (**d**), the octoploid genome (32 genes) fractionates again to yield 16 genes (**e**), which duplicate (to 32 genes) in yet another WGD (**f**), after which fractionation yields 16 genes in the 16-ploid (**g**). The resulting number of genes is the same as in the fractionated genome resulting from the first WGD (**c**), with only 6 more genes than the original diploid ancestor (**a**). **h**, The resulting genome after intergenic DNA contraction at any stage (**a**–**g**) has thus survived a high deletion rate via the net accrual of very few gene duplicates following sequential WGDs. *U. gibba* has in fact fractionated down to single copy two-thirds of its genes syntenic to tomato genes since its three WGDs. **i**, An interplay of deletion and retroelement proliferation rates relates to a continuum of plant genome size evolution, with WGDs providing short-term buffering against loss of crucial gene functions in small genomes affected by high endogenous deletion rates. Small genomes result when the recombinational deletion rate is high relative to retroelement proliferation and WGD, vice versa with large genomes.

Taken together, we infer from our analyses of *U. gibba* coding sequence that natural selection preserved a core set of gene functions, most of which have returned to single copy along with considerable genomic fractionation after three WGDs. Relaxed selection pressure for unnecessary functions probably led to gene losses, whereas in other cases, gene family expansions may have been promoted by selection. Evidence for localized selection on the *U. gibba* gene complement, however, does not provide support for the existence of genome-wide selective forces that might favour reduction of nonessential, non-coding DNA.

It has been argued that increased mutation pressure can enhance natural selection against non-essential DNA[14]. We proposed previously that enhanced molecular evolutionary rates caused by mutagens could have made the *U. gibba* genome more susceptible to natural selection[3,15]. This could now be evaluated, because information on the mutational diversity (θ) stored within a single genome is retrievable. θ, when small as in *Arabidopsis*[16], closely approximates heterozygosity. We found that *U. gibba* does not have a θ value substantially different from that of *Arabidopsis* (Supplementary Information 6). As such, it is possible that the population genetic environment underlying *U. gibba* genome evolution did not engender special sensitivity to natural selection beyond that experienced by *Arabidopsis* with its larger proportion of non-coding DNA.

Collectively, our analyses highlighting total gene complement, sequential WGD and mutational diversity estimates for *U. gibba* raise quandaries regarding the evolution of its contracted genome. It is possible that inherent molecular mechanisms favouring deletion dominated nuclear genome size reduction in a population genomic background where selection was too weak to counteract such a burden. Some intrinsic molecular biases are known to correlate with genome size differences. For example, the net DNA deletion bias caused by double-strand break repair in *Arabidopsis* (120 Mb[7]) is greater than that of tobacco (5.1 gigabases (Gb)[17]), and deletions are larger as well[18]. A similar bias occurs in *Arabidopsis thaliana* compared to its larger-genome relative *Arabidopsis lyrata*[9]. Biased gene conversion, which is associated with (G+C)-rich sequences such as those found throughout the *U. gibba* genome[19], leads to its own inherent deletion bias[20], which has been argued to be an important neutral process behind other genome size reductions[21]. Of course, a molecular-mechanistic deletion bias does not preclude that selection still enhances fixation of such deletions.

Regarding a potential role of polyploidy in genome contraction, we propose that for small genomes facing a strong internal deletion bias, WGDs, by the creation of duplicates throughout the genome, might transiently buffer against loss of essential genes (Fig. 3). Interestingly, phylogenetic evidence indicates that genome evolution is highly dynamic in Lentibulariaceae, with nuclear DNA contents ranging from 60 Mb to 1.5 Gb[22]. Sequencing of additional Lentibulariaceae genomes is warranted to ascertain the basis for these differences. Moreover, because molecular dating analyses place the divergence of *Utricularia* from its carnivorous relative *Pinguicula* at approximately 40 million years before present (Myr BP)[23], and that of *U. gibba* from other *Utricularia* species as recently as 5–15 Myr BP (Supplementary Information section 9), additional high-quality Lentibulariaceae genomes should permit phylogenetic dating of the sequential WGD events that occurred after common ancestry with tomato, approximately 87 Myr BP[23].

In summary, *U. gibba* genome architecture demonstrates that angiosperms can evolve diverse gene landscapes while overall genome size contracts, not only during expansions. Furthermore, in contrast to recent publications that highlight a crucial functional role of non-coding DNA in complex organisms such as animals[24], the necessary genomic context required to make a flowering plant may not require substantial hidden regulators in the non-coding 'dark matter' of the genome.

METHODS SUMMARY

Genomic DNA from *U. gibba* was subjected to a hybrid 454, Illumina and Sanger sequencing strategy. Approximately 5.2 Gb of sequence data were generated, consisting of 1.9 Gb of shotgun reads, 1.5 Gb of mate-pair reads, 1.5 Gb of paired-end reads and 119.5 Mb of Sanger reads; these were assembled using Newbler version 2.6. The assembly was filtered for organellar and environmental DNA, and validated by primer walking of representative scaffolds and random fosmid sequencing (Supplementary Information sections 1.4–1.6). A transcriptome from pooled plant parts served as a gene prediction and annotation aid (Supplementary Information section 2.3). Transposable elements were identified using the REPET package (Supplementary Information section 2.1). Non-coding RNAs were identified using tRNAscan-SE, RNAMMER, snoscan, and SRPscan (Supplementary Information sections 2.2 and 4). Gene models were predicted using AUGUSTUS with a transcriptome-derived training set (Supplementary Information section 2.3.2). Synteny to other plant genomes was analysed using CoGe (Supplementary Information section 7). Frequencies of Pfam domains among gene models, and their significant differences, were calculated for *U. gibba* and several other plant genomes (Supplementary Information section 2.4). Gene models from *U. gibba* and other plant species were clustered into orthogroups using OrthoMCL, annotated using Blast2GO, and studied for expansions and contractions of gene memberships (Supplementary Information section 2.5). Selected gene families from *U. gibba*, *Arabidopsis* and tomato were subjected to phylogenetic analysis (Supplementary Information section 2.5.2). The *U. gibba* genome was scanned for single-copy genes identified from other plant genomes (Supplementary Information section 2.6). Promoters and untranslated regions (UTRs) were studied *in silico*, selected UTRs were amplified by PCR and sequenced, and selected promoters were analysed *in vivo* using transient expression assays (Supplementary Information section 3). Genome compositional features were compared to *Arabidopsis* (Supplementary Information section 5). Population genomic parameters were calculated using the PSMC and mlRho applications (Supplementary Information section 6). Organelle genomes were assembled using Newbler version 2.6 and Megamerger, and annotated using

DOGMA (Supplementary Information section 8). Molecular evolutionary rates and divergence times were estimated using BEAST and HyPhy (Supplementary Information section 9).

Full Methods and any associated references are available in the online version of the paper.

Received 14 December 2012; accepted 25 March 2013.
Published online 12 May 2013.

1. Bennetzen, J. L. & Kellogg, E. A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9,** 1509–1514 (1997).
2. Taylor, P. *The Genus Utricularia: a Taxonomic Monograph* (Kew Publishing, 1989).
3. Albert, V. A., Jobson, R. W., Michael, T. P. & Taylor, D. J. The carnivorous bladderwort (*Utricularia*, Lentibulariaceae): a system inflates. *J. Exp. Bot.* **61,** 5–9 (2010).
4. Płachno, B. J. & Swiatek, P. Unusual embryo structure in viviparous *Utricularia nelumbifolia*, with remarks on embryo evolution in genus *Utricularia*. *Protoplasma* **239,** 69–80 (2010).
5. Duarte, J. M. *et al.* Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10,** 61 (2010).
6. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
7. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
8. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485,** 635–641 (2012).
9. Hu, T. T. *et al.* The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genet.* **43,** 476–481 (2011).
10. Woodhouse, M. R. *et al.* Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS Biol.* **8,** e1000409 (2010).
11. Devos, K. M., Brown, J. K. & Bennetzen, J. L. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Res.* **12,** 1075–1079 (2002).
12. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
13. Kibriya, S. & Jones, J. I. Nutrient availability and the carnivorous habit in *Utricularia vulgaris*. *Freshw. Biol.* **52,** 500–509 (2007).
14. Lynch, M., Koskella, B. & Schaack, S. Mutation pressure and the evolution of organelle genomic architecture. *Science* **311,** 1727–1730 (2006).
15. Ibarra-Laclette, E. *et al.* Transcriptomics and molecular evolutionary rate analysis of the bladderwort (*Utricularia*), a carnivorous plant with a minimal genome. *BMC Plant Biol.* **11,** 101 (2011).
16. Nordborg, M. *et al.* The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3,** e196 (2005).
17. Leitch, I. J. *et al.* The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann. Bot.* **101,** 805–814 (2008).
18. Kirik, A., Salomon, S. & Puchta, H. Species-specific double-strand break repair and genome evolution in plants. *EMBO J.* **19,** 5562–5566 (2000).
19. Ibarra-Laclette, E., Albert, V. A., Herrera-Estrella, A. & Herrera-Estrella, L. Is GC bias in the nuclear genome of the carnivorous plant *Utricularia* driven by ROS-based mutation and biased gene conversion? *Plant Signal. Behav.* **6,** 1631–1634 (2011).
20. Assis, R. & Kondrashov, A. S. A strong deletion bias in nonallelic gene conversion. *PLoS Genet.* **8,** e1002508 (2012).
21. Nam, K. & Ellegren, H. Recombination drives vertebrate genome contraction. *PLoS Genet.* **8,** e1002680 (2012).
22. Greilhuber, J. *et al.* Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size. *Plant Biol.* **8,** 770–777 (2006).
23. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97,** 1296–1303 (2010).
24. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
25. Aagaard, J. E., Olmstead, R. G., Willis, J. H. & Phillips, P. C. Duplication of floral regulatory genes in the Lamiales. *Am. J. Bot.* **92,** 1284–1293 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank I. Bravo-Carmona for growing plants for this research. We also thank T. Mailund for discussion and critical comments on the manuscript, M. Stanke and K. Hoff for their assistance in the use of AUGUSTUS software, and P. Guzmán-Villate for providing the cell suspension used for transient expression assays. Special thanks goes to the members of the 'laboratorio de servicios genómicos' of LANGEBIO, CINVESTAV for sequencing services and their help. This research was supported by CONACyT (Mexico) via general support to LANGEBIO, HHMI grant 4367 (to L.H.-E.), the College of Arts and Sciences, University at Buffalo (to V.A.A.), and the NSF (0922742 to V.A.A.). E.I.-L. is indebted to CONACyT (Mexico) for a PhD fellowship. We acknowledge the US Department of Energy Joint Genome Institute for *Mimulus* data (available at http://www.phytozome.net/mimulus).

Author Contributions E.I.-L., V.A.A. and L.H.-E. conceived of and led the study. E.I.-L., V.A.A. and L.H.-E. wrote the paper with significant contributions by E.L., L.C.-P. and A.J.W.; E.I.-L., G.H.-G., C.A.P.-T., T.-H.C., T.L., M.J.A.J., S.C.S., A.O.-A., S.A.C.-P. and M.d.J.O.-E. collected data. E.I.-L., E.L., L.C.-P., T.-H.C., T.L., A.J.W., M.A.-V., E.G.-C., G.A.-H., H.H., A.E.M., S.X., M.L. and V.A.A. analysed data. J.S., T.P.M., T.M., D.B. and A.H.-E. provided materials. A.F.-C. and J.I.C.-L. provided bioinformatic support. All authors read and approved the final manuscript.

Author Information Files containing raw sequence reads and quality scores were deposited in the Sequence Read Archive of the National Center for Biotechnology Information (NCBI). Primary accession numbers: SRS399135 (454 reads), SRS399163 (MiSeq reads), SRS399167 (fosmid Ion Torrent reads) and SRS399168 (RNAseq Ion Torrent reads). The *U. gibba* genome assembly and gene models are available on CoGe (http://genomevolution.org/CoGe/). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to V.A.A. (vaalbert@buffalo.edu) or L.H.-E. (lherrera@langebio.cinvestav.mx).

METHODS

Utricularia gibba was collected in the Umécuaro municipality, Michoacán, Mexico. For flow cytometry analysis, nuclei were isolated from shoot-like structures and flowers, stained with 1.5 ml 4′,6-diamidino-2-phenylindole, and their fluorescence measured after ultraviolet excitation. *Arabidopsis thaliana* was used as an internal standard to calculate *U. gibba* nuclear DNA content. The genome size estimated was 77.38 Mb.

Nuclear DNA was isolated from *U. gibba* shoot-like structures, then amplified and sheared to obtain DNA fragments ranked according to the sizes required for sequencing libraries (1 kb, 2 kb, 2–4 kb or 7–9 kb). For whole-genome shotgun sequencing, four distinct shotgun libraries (one 3 kb and three 8 kb mate-pair libraries) were constructed. Preparation, amplification and sequencing of these libraries were performed using Roche GS FLX Titanium Sequencing Kits and Genome Sequencer FLX Instruments following the manufacturer's protocols. One additional shotgun library was constructed and sequenced using the GS FLX XL+ Sequencing kit and corresponding platform. Additionally, one paired-end library of ~450 bp was prepared using Illumina's paired-end kit. The nuclear DNA was sheared with a Covaris S2 ultrasonicator and the library was sequenced (twice) as 2×250 bp on an Illumina MiSeq. Finally, conventional Sanger reads were generated with an ABI 3730xl sequencer using the Big Dye–terminator Cycle Sequencing kit. Recombinant clones (pJET1.2/blunt Cloning Vector) were used to transform DH10b cells to obtain two genomic libraries ((1) 43,968 clones, average insert size 1.2 kb, and (2) 55,680 clones, average insert size 4 kb), and clones were sequenced both uni- and bidirectionally. In total, ~5.2 Gb of sequence data was generated, consisting of 1.9 Gb of shotgun reads, 1.5 Gb of mate-pair reads, 1.5 Gb of paired-end reads and 119.5 Mb of Sanger reads (Supplementary Table 2).

The 454, Sanger and MiSeq reads were assembled using Newbler version 2.6 *de novo* genome assembler (with the -scaffold option). Vector and poor quality regions were masked in the Sanger reads using the LUCY2 software. Natural and artificial duplicates in pyrosequencing reads were eliminated using the CD-HIT pipeline. The MiSeq read pairs (2×250) were merged and adaptor-trimmed with SeqPrep using default settings. Paired-end reads that did not overlap with at least 10 bases were subjected to stringent read filtering and trimming before assembly. Reads were trimmed with a sliding window approach (window size 10 bases, shift 1 base). Illumina bases were kept until the average quality score Q of 10 adjacent bases was below $Q = 25$. Reads were removed if they were shorter than 30 bases after trimming, had at least one uncalled base, contained the adaptor sequence, or had less than two-thirds of the bases of the first half of the read with quality values of $Q \geq 30$. Orphan reads were discarded to keep pairs only. Redundant read pairs that may originate from PCR artefacts were also removed by comparing the sequences of the read pairs. Out of 6,215,172 read pairs, 28% could be merged and 60% passed the stringent filtering. The average length of the merged reads was 459 bp. The filtered MiSeq pairs were exclusively used for scaffolding by trimming them to 49 bases. We generated a total of 4.7 billion high-quality base pairs from 20.3 million high-quality reads. After *de novo* assembly, contaminating sequences from organellar and environmental DNA were removed by a GC value and coverage-based filtering process. The *U. gibba* assembly spanned, with around 35-fold genome coverage, 81.87 Mb including embedded gaps (N50 = 80,839 bp, the weighted mean statistic such that 50% of the assembly is contained in contigs and scaffolds equal to or larger than this value). The total length of the assembled genome was about 5.73% greater than the genome size estimated by flow cytometry of isolated nuclei stained with DAPI (77.38 Mb).

Our assembly of the *U. gibba* genome was verified by single-pass primer walking resequencing of a ~100 kb window (total) from two randomly selected scaffolds. Additionally, using the pCC1FOS vector, a fosmid library with ~1,000 clones was generated. The complete sequences of 53 end-sequenced fosmids (with BLAST hits to the *U. gibba* genome) were obtained with an estimated coverage of ~250×. The complete alignments of fosmid sequences to the *U. gibba* whole genome sequence revealed that we were able to generate a shotgun assembly with only limited potential misassemblies.

Transposable elements in the *U. gibba* genome were identified both at the DNA and protein level. The REPET package was used to search for transposable elements within the *U. gibba* genome. To confirm the degree of completeness of *U. gibba* LTR retrotransposons, characteristic elements (both 5′- and 3′-long terminal repeats (LTRs), primer binding sites (PBSs), polypurine tracts (PPTs), and conserved protein domains and their positions) were identified using the LTR-Finder program. We took a computational approach to gain insight into the different RNA-mediated gene regulatory pathways present in *U. gibba*. Non-coding RNAs (ncRNAs), including miRNAs, small nuclear RNAs, tRNAs, ribosomal RNAs and H/ACA-box small nucleolar RNAs, were identified using INFERNAL software by searching against the Rfam database.

For transcriptome sequencing, total RNA was extracted from whole plants, shoot-like structures, inflorescences and traps using TRIzol according to the manufacturer's instructions. To represent all *U. gibba* organs, 2 μg of RNA from each sample were pooled. cDNA synthesis was performed as described previously. The sequences were assembled with Newbler version 2.6

The AUGUSTUS program was trained on the *U. gibba* genome using 37,799 Isotig sequences. First, using the $\text{AUGUSTUS}_{\text{beta}}$ web server training tool and the *U. gibba* genome and transcriptome sequences (Isotigs), a data set with training gene structures was generated. Using this training set, parameters required by AUGUSTUS were calculated. Gene models in the *U. gibba* genome sequence were predicted *ab initio* as well as with hints, running AUGUSTUS locally with newly optimized parameters.

To analyse the distribution of gene families over different plant species, we identified the Pfam domains present from gene models predicted in the *Arabidopsis*, tomato, grape, *Mimulus* and papaya genomes. To compare the abundance of domains in proteins of different plant species we used a likelihood ratio test method (see Supplementary Information for more details). Clustering of homologous genes for the *U. gibba*, *Arabidopsis*, tomato, grape and papaya genomes was performed using OrthoMCL on the predicted protein sequences of all the five genomes. All *U. gibba* gene models were processed through the Blast2GO program to assign functions. We closely surveyed the first 100 OrthoMCL clusters showing *U. gibba* gene family member expansions, and then the first 100 showing contractions. We performed detailed phylogenetic classifications of five well-known transcription factor families (MADS, TCP, GRAS, ARF and AUX/IAA) using maximum likelihood and neighbour joining methods to provide highly focused views of gene family expansion and contraction in *U. gibba* relative to *Arabidopsis* and tomato. Using bidirectional best BLAST and synteny analysis (SynMap within CoGe), we calculated the proportions of previously reported single-copy genes (in *Arabidopsis*, *Vitis*, poplar and rice) that are also present as single copy in the *U. gibba* genome.

We estimated the average length of intergenic regions considering pairs of adjacent genes as either convergent (→ ←), divergent (← →), or tandem (→ → or ← ←). A total of 14 adjacent gene pairs (5 convergent, 4 divergent and 5 tandem) were selected to estimate UTR sizes in the *U. gibba* genome by random amplification of cDNA ends (RACE-PCR). For a *rbcS* gene promoter from *U. gibba*, we identified and studied the compaction of the I- and G-boxes and two other motifs almost always conserved in other species. The functionality of some promoters in *U. gibba* was tested by transient expression assay.

We applied the pairwise sequentially Markovian coalescent (PSMC) model, which was originally applied to human and other mammalian genomes, to study the mutational diversity of the *U. gibba* genome and effective population size (N_e) over time. The *Arabidopsis thaliana* genome (and reads from accession SRX158512) was treated similarly. In PSMC coalescent simulations, N_e is inferred from heterozygosity of the sequenced genome ($\theta = 4N_e\mu$). The mlRho application was similarly used to estimate genome-wide and window-based (100 kb, 75 kb, 50 kb and 25 kb) θ values.

For analyses of whole genome duplications, we focused on comparing the genomes of *Solanum lycopersicum* and *U. gibba* using the SynMap tool in the online CoGe portal (http://genomevolution.org/CoGe/). CoGe contains two major applications to help evaluate and estimate syntenic depth: SynMap and SynFind. We compared tomato to *U. gibba* using two parameter sets that differ in the window size of genes used to define a minimum number of colinear genes allowing two regions to be called syntenic. Fractionation depth refers to the number of syntenic genes that reduce to single-, double- or *n*-copy over the course of *U. gibba*'s three independent WGDs since common ancestry with tomato. Results were generated from SynMap via a master table of all genes in tomato along with their matching syntenic regions in *U. gibba*. GEvo microsyntenic analyses were performed on selected regions determined to be syntenic using SynMap and SynFind.

Scaffolds/contigs originating from the plastid and mitochondrial genomes of *U. gibba* were identified during the process of *de novo* assembly using Newbler version 2.6. These were further assembled and annotated using the Megamerger program and DOGMA web tool.

In order to investigate the divergence time of *U. gibba* from other *Utricularia* species, we obtained phylogenetic data sets for the family Lentibulariaceae from three regions of the chloroplast genome and one region of the mitochondrial genome. We applied the BEAST program to estimate divergence dates, and both this program and HyPhy to study molecular evolutionary rates.

LETTER

doi:10.1038/nature12198

Gut metagenome in European women with normal, impaired and diabetic glucose control

Fredrik H. Karlsson[1]*, Valentina Tremaroli[2]*, Intawat Nookaew[1], Göran Bergström[2], Carl Johan Behre[2], Björn Fagerberg[2], Jens Nielsen[1] & Fredrik Bäckhed[2,3]

Type 2 diabetes (T2D) is a result of complex gene–environment interactions, and several risk factors have been identified, including age, family history, diet, sedentary lifestyle and obesity. Statistical models that combine known risk factors for T2D can partly identify individuals at high risk of developing the disease. However, these studies have so far indicated that human genetics contributes little to the models, whereas socio-demographic and environmental factors have greater influence[1]. Recent evidence suggests the importance of the gut microbiota as an environmental factor, and an altered gut microbiota has been linked to metabolic diseases including obesity[2,3], diabetes[4] and cardiovascular disease[5]. Here we use shotgun sequencing to characterize the faecal metagenome of 145 European women with normal, impaired or diabetic glucose control. We observe compositional and functional alterations in the metagenomes of women with T2D, and develop a mathematical model based on metagenomic profiles that identified T2D with high accuracy. We applied this model to women with impaired glucose tolerance, and show that it can identify women who have a diabetes-like metabolism. Furthermore, glucose control and medication were unlikely to have major confounding effects. We also applied our model to a recently described Chinese cohort[4] and show that the discriminant metagenomic markers for T2D differ between the European and Chinese cohorts. Therefore, metagenomic predictive tools for T2D should be specific for the age and geographical location of the populations studied.

The composition of the gut microbiota differs among geographical locations, and between elderly people, in whom T2D incidence is high, and younger subjects[6–10]. In addition, studies of T2D are complicated by the heterogeneous manifestations and mixed aetiology of the disease, and confounded by the effects of age, gender, degree of glucose control and concomitant treatment. In this study, we examined the composition and function of the faecal microbiota in a well-characterized population of 70-year-old European women to minimize sources of variation. Our cohort was selected using a stratified randomized method from a population-based screening sample[11,12] and classified into three similarly sized subgroups: women who had T2D ($n = 53$), impaired glucose tolerance (IGT; $n = 49$) or normal glucose tolerance (NGT; $n = 43$) (Methods and Supplementary Tables 1–3). Genomic DNA was extracted from faecal samples using a standard procedure[13] and sequenced on Illumina HiSeq 2000. In total, we obtained 453 gigabases (Gb) of paired-end reads, with an average of 3.1 ± 1.8 Gb (mean $\pm$ s.d.) for each sample (Supplementary Table 4).

To determine the composition of the gut microbiota, we aligned filtered Illumina reads to 2,382 non-redundant reference genomes obtained from the NCBI and HMP databases (http://www.hmpdacc.org) (Supplementary Table 5) using our recently published MEDUSA platform[5]. We compared the composition of T2D and NGT communities and observed increases in the abundance of four *Lactobacillus* species and decreases in the abundance of five *Clostridium* species in the T2D group (adjusted $P < 0.05$, Wilcoxon rank sum test) (Supplementary Fig. 1a and Supplementary Table 6). In the total cohort, *Lactobacillus* species correlated positively with fasting glucose and HbA1c (glycosylated haemoglobin), a long-term measure of blood glucose control (adjusted $P < 0.05$, Spearman correlation). By contrast, *Clostridium* species correlated negatively with fasting glucose, HbA1c, insulin, C-peptide and plasma triglycerides, and positively with adiponectin and HDL (Supplementary Fig. 1b and Supplementary Table 7). These correlations are relevant for T2D because high triglycerides and low HDL levels are components of the dyslipidaemia typically found in T2D, whereas serum levels of the insulin-sensitizing hormone adiponectin are reduced in people at risk of T2D (ref. 14). Importantly, these *Lactobacillus* and *Clostridium* species did not correlate with body mass index (BMI), waist circumference or waist-to-hip ratio (WHR) (Supplementary Fig. 1b).

To identify microbial species independently of reference genomes and fully exploit the information contained in the metagenomic data, we performed *de novo* assembly of filtered sequence data. The total length of the assembly was 13.6 Gb, from which 18.6 million genes with a length longer than 100 base pairs (bp) could be predicted. We created a non-redundant gene catalogue for our cohort and merged it with the MetaHIT gene catalogue[15]. The merged gene catalogue was used to align reads. The faecal microbiota of NGT, IGT and T2D women contained similar numbers of genes (Supplementary Fig. 2). We clustered these genes based on their profile across samples with the assumption that genes from the same genome should have a similar abundance within each subject. We considered only genes that were shared among at least 10 subjects (2.9 million genes) and calculated the correlation coefficient across subjects. We clustered sets of genes with high correlation between them (Pearson rho > 0.85) and defined these sets as metagenomic clusters (MGCs) (Supplementary Fig. 3). The 800 largest MGCs contained at least 104 genes, and 550,188 genes were included in total (Supplementary Table 8; distribution of the number of genes in MGCs shown in Fig. 1a).

To determine the phylogenetic origin of the MGCs, we blasted the genes in each cluster against the NCBI non-redundant catalogue and determined the lowest common ancestor (LCA) by requiring that at least 50% of the genes had a best hit to the same phylogenetic group (Supplementary Fig. 4). This analysis showed that 36% of the MGCs had an LCA at the species level (Fig. 1b), and that MGCs with an LCA at the order level (30%) were mainly Clostridiales (98%) and few Bacteroidales (2%). The Clostridiales order is very diverse and reference genomes might be lacking in public databases, thus explaining the difficulty of the taxonomic characterization.

We tested the abundance of the 800 largest MGCs in NGT and T2D samples, and found 26 clusters to be differentially abundant between the two groups (adjusted $P < 0.05$, Wilcoxon rank sum test) (Fig. 1c and Supplementary Table 9). The MGCs most significantly enriched in

[1]Department of Chemical and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [2]The Wallenberg Laboratory and Sahlgrenska Center for Cardiovascular and Metabolic Research, Department of Molecular and Clinical Medicine, Institute of Medicine, University of Gothenburg, SE-413 45 Gothenburg, Sweden. [3]Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Receptology and Enteroendocrinology, Faculty of Health Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.
*These authors contributed equally to this work.

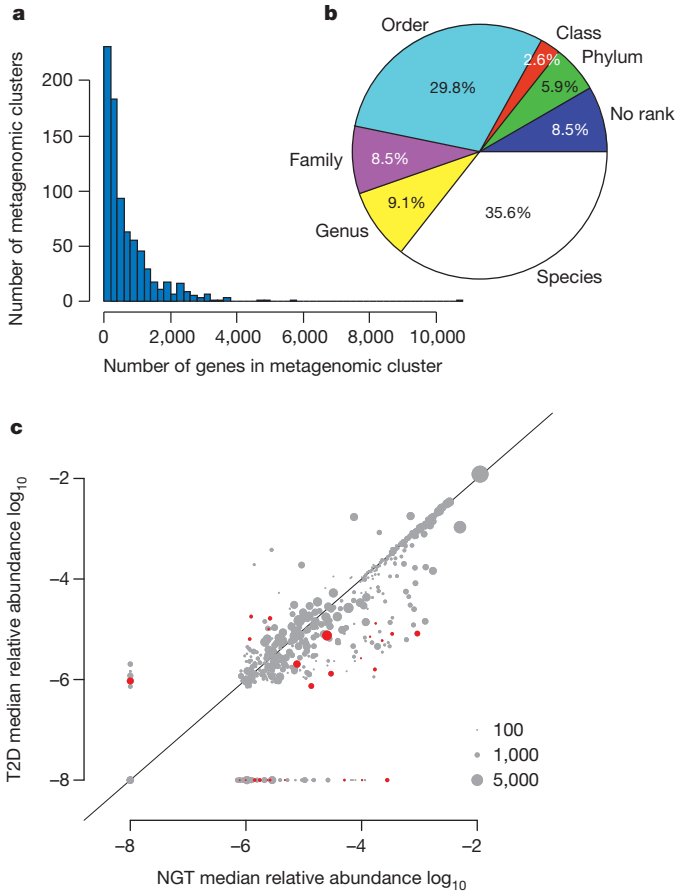


Figure 1 | Definition of MGCs and identification of differentially abundant MGCs in T2D and NGT. **a**, Histogram of the number of genes in the 800 largest MGCs, all with more than 100 genes. **b**, Pie chart of the taxonomic annotation level of MGCs. **c**, Scatter plot of median MGC abundance in T2D ($n = 53$) and NGT ($n = 43$) women. Grey points represent MGCs not differentially abundant between groups, and red points represent differentially abundant MGCs (adjusted $P < 0.05$, Wilcoxon rank sum test).

T2D women were a Clostridiales identified at order level and two *Clostridium clostridioforme*. Two other MGCs were enriched in T2D microbiota, and were identified at species levels as *Lactobacillus gasseri* and *Streptococcus mutans*. *C. clostridioforme* correlated positively with triglyceride and C-peptide levels, whereas *L. gasseri* correlated positively with fasting glucose and HbA1c (Fig. 2 and Supplementary Table 10). Twenty-one MGCs were significantly depleted in T2D, including *Roseburia* (that is, *Roseburia*_272), two unknown *Clostridium* species, several Clostridiales, two *Eubacterium eligens*, Coriobacteriaceae and one *Bacteroides intestinalis*. In the total cohort, the clostridial MGCs correlated negatively with C-peptide, insulin and triglyceride levels, whereas *B. intestinalis* correlated negatively with insulin and waist circumference (Fig. 2 and Supplementary Table 10). These results largely agree with those obtained from the species-based analyses (Supplementary Fig. 1).

To test whether the microbiota composition can identify diabetes status, we trained a random forest model in a training set of the NGT and T2D subjects using the profiles of species and MGCs. We evaluated its performance using a tenfold cross-validation approach and scored the predictive power in a receiver operating characteristic (ROC) analysis. The discriminatory power of species and MGCs was calculated as the area under the ROC curve (AUC). T2D was identified more accurately with MGCs (highest AUC = 0.83) than with microbial species (highest AUC = 0.71) (Fig. 3a and Supplementary Table 11). The increased AUC for the MGC-based model can be explained by the fact that MGCs also provide taxonomical and functional information for unknown species. Therefore, the MGC-based method has the advantage that it can also be applied when reference genomes are missing. When BMI, WHR and waist circumference were used for predicting T2D, we obtained a maximum AUC of 0.70 for waist circumference (BMI, AUC = 0.58; WHR, AUC = 0.60), thus showing that the composition of the microbiota determined by MGCs correlates better with diabetes than these known T2D risk factors[16]. Importantly, the T2D score obtained based on MGCs is similar to other published scores that combine several known risk factors for diabetes development (for example, the FINDRISC score, validated in several countries[1]).

L. gasseri had the highest score for the identification of T2D women in both models (species and MGCs; Fig. 3b, c). Lactobacilli and clostridia were among the ten most important bacteria in the species model (Fig. 3b), whereas *Roseburia*, several Clostridiales, *B. intestinalis*, *C. clostridioforme* and Coriobacteriaceae were among the ten most important clusters in the model based on MGCs (Fig. 3c). The two models indicated different bacterial groups as most discriminant for T2D identification, but the bacteria identified by the MGC model had higher scores than those identified by the species model (Fig. 3b, c). Notably, the MGC model identified *Roseburia* and *Faecalibacterium prausnitzii* as highly discriminant for T2D. These bacteria are known human gut colonizers and butyrate producers[17], and have been linked to improved insulin sensitivity and diabetes amelioration in studies of the human faecal microbiota[18,19].

Many patients are not diagnosed with T2D until cardiovascular complications are apparent[20], but IGT and other metabolic defects often appear before T2D develops[21]. We used our random forest model trained for the discrimination of NGT and T2D individuals to stratify the 49 IGT women of the cohort: 10 IGT women were included in the NGT subgroup, whereas 34 were included in the T2D subgroup (5 could not be classified, as the probability of being either NGT or T2D was 0.5 ± 0.02; Fig. 4a). The characteristics of the two subgroups stratified according to the faecal metagenome showed that plasma levels of triglycerides and C-peptide were significantly higher in the subgroup identified as T2D than in the subgroup identified as NGT (Fig. 4b, c).

To characterize microbial functions, we annotated all of the genes in our catalogue to the KEGG database (version 59). We used the reporter feature algorithm[22,23] in combination with the KEGG metabolic network, pathway annotations and the information about relative gene abundance to identify reporter pathways (that is, pathways with significantly differentially abundant KEGG orthologues) that were associated with T2D and NGT status. We found that NGT and T2D communities had different functional composition and several reporter pathways were differentially abundant in T2D and NGT women (Supplementary Table 12). The pathways that showed the highest scores for enrichment in T2D metagenomes included KEGG orthologues for starch and glucose metabolism (39 out of 46), fructose and mannose metabolism (37 out of 49), and ABC transporters for amino acids, ions and simple sugars (123 out of 174) (Supplementary Table 12). These results are in agreement with previous studies showing an increase in microbial functions for energy metabolism and harvest in the obese microbiome[2,3]. Other metabolic pathways containing KEGG orthologues enriched in women with T2D included glycerolipid metabolism and fatty acid biosynthesis. Pathways for cysteine and methionine metabolism were also enriched in T2D; these pathways are related to glutathione synthesis and may be important for response to oxidative stress.

Similar to our observations, genes related to membrane transporters and oxidative stress resistance were enriched also in the Chinese T2D metagenome[4]. In our study, microbial functions enriched in NGT women were related to flagellar assembly and riboflavin metabolism (Supplementary Table 12). Interestingly, the metagenome of healthy individuals in the Chinese cohort was also enriched in functions related to flagellar assembly and metabolism of cofactors and vitamins[4].

We next investigated whether the composition and functionality of the gut microbiota is influenced by factors other than prevalent T2D,

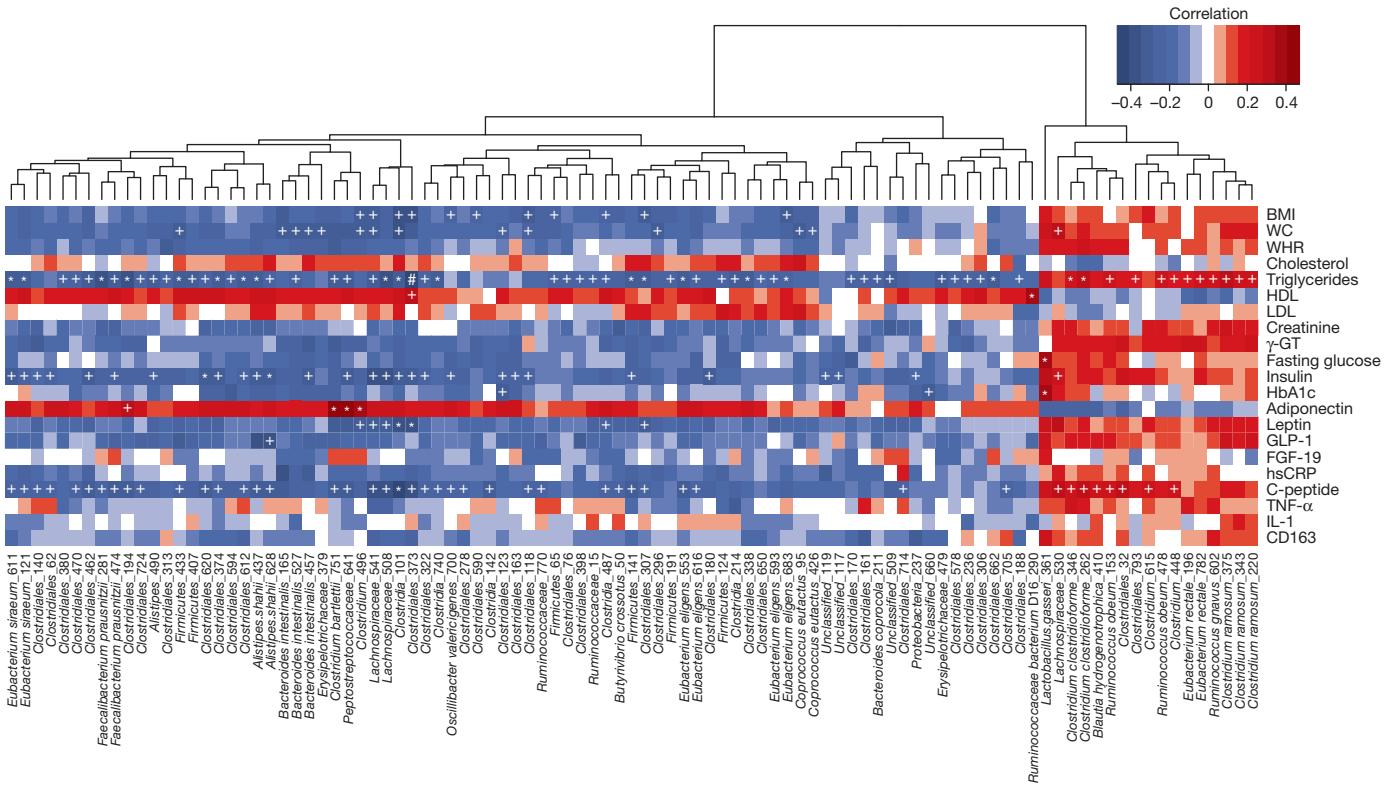


Figure 2 | Associations of MGCs with clinical biomarkers. Spearman's rank correlation coefficients and *P* values for the correlations are listed in Supplementary Table 10. $n = 145$; '+' denotes adjusted $P < 0.05$; '*' denotes adjusted $P < 0.01$; '#' denotes adjusted $P < 0.001$. FGF-19, fibroblast growth factor 19; γ-GT, γ-glutamyltransferase; GLP-1, glucagon-like peptide 1; HDL, high-density lipoprotein; hsCRP, high-sensitivity C-reactive protein; LDL, low-density lipoprotein; WC, waist circumference.

such as family history of diabetes, medication (that is, statins and metformin) or degree of blood glucose control (as measured by HbA1c levels; <47 mmol mol^{-1} indicates good control). We observed that several microbial species and gene functions were associated with metformin and glucose control (Supplementary Results, Supplementary Fig. 5a, b and Supplementary Tables 13–16). However, these associations should not have a major confounding effect on the model for the discrimination of T2D women based on faecal microbiota composition as only two of the species included in our model were affected by the use of metformin (*Clostridium botulinum* B str. Eklund 17B and *Clostridium* sp. 7_2_43FAA; Supplementary Table 13) and two others were affected by poor glucose control (*Clostridium thermocellum* DSM 1313 and *Streptococcus* sp. C150; Supplementary Table 14). Furthermore, these associations were not identified using an MGC-based approach (Supplementary Fig. 5c, d).

The similarities between the results obtained in our study and those reported previously[4] clearly underline the link between gut microbiota and T2D. To confirm further the similarities independent of methodological differences, we analysed the Chinese metagenomic data with our bioinformatics platform and compared the results with those from our cohort (Supplementary Results, Supplementary Figs 6–15 and Supplementary Tables 17–19).

The Chinese and European populations clustered separately in principal component analysis plots of species and MGCs abundance (Supplementary Fig. 10), which may be a result of different genetics and/or dietary habits. However, in agreement with the previous study[4] we observed that *Clostridium clostridioforme* MGCs were increased whereas *Roseburia*_272 was decreased in T2D metagenomes from both cohorts (Supplementary Tables 9 and 18). *C. clostridioforme* is a mixture of three opportunistic pathogens (*C. bolteae*, *C. hathewayi* and *C. clostridioforme*) that have been associated with bacteraemia and infections in humans[24], whereas *Roseburia* contains gut bacteria able to produce butyrate. Gut microbiota transplantations from lean donors to recipients with metabolic syndrome have recently been shown to increase *Roseburia* and butyrate levels together with improved insulin sensitivity[18], thus suggesting the importance of butyrate-producing bacteria for blood glucose regulation in humans.

We also found increased *Lactobacillus* species and MGCs in both T2D cohorts (Supplementary Tables 6, 9, 17 and 18). Increased *Lactobacillus* levels in T2D patients were also observed in another small study[25], which used 16S ribosomal DNA pyrosequencing to analyse the microbiota composition of T2D patients and healthy men. A positive correlation between *Lactobacillus* abundance and blood glucose levels was shown[25], in agreement with our study (Fig. 2 and Supplementary Fig. 1b). The increase in *Lactobacillus* could be a consequence of increased glucose levels in the intestine, as increased lactobacilli resulting from increased salivary glucose have been measured in children with insulin-dependent diabetes mellitus[26].

We used the MGCs identified in our study to train a new random forest model on Chinese metagenomes, and used this model to classify Chinese subjects into T2D and controls. We observed an AUC of 0.82 (Supplementary Fig. 14 and Supplementary Table 19), which is in line with the value of 0.81 reported previously[4] and similar to the value reported for the classification of NGT and T2D women in our cohort (0.83; Fig. 3a and Supplementary Table 11). We observed that the most discriminatory MGCs differed between the Chinese subjects and our cohort (Fig. 3b, c and Supplementary Fig. 15). In particular, *Akkermansia* did not contribute to the classification of T2D women in our cohort, whereas *Lactobacillus* did not contribute to the classification of T2D patients in the Chinese population, thus suggesting that classifiers for T2D based on species are population specific. However, it should be noted that, in contrast to our homogenous cohort (70-year-old women), the T2D population in the previous study was older and included more men than the control population,

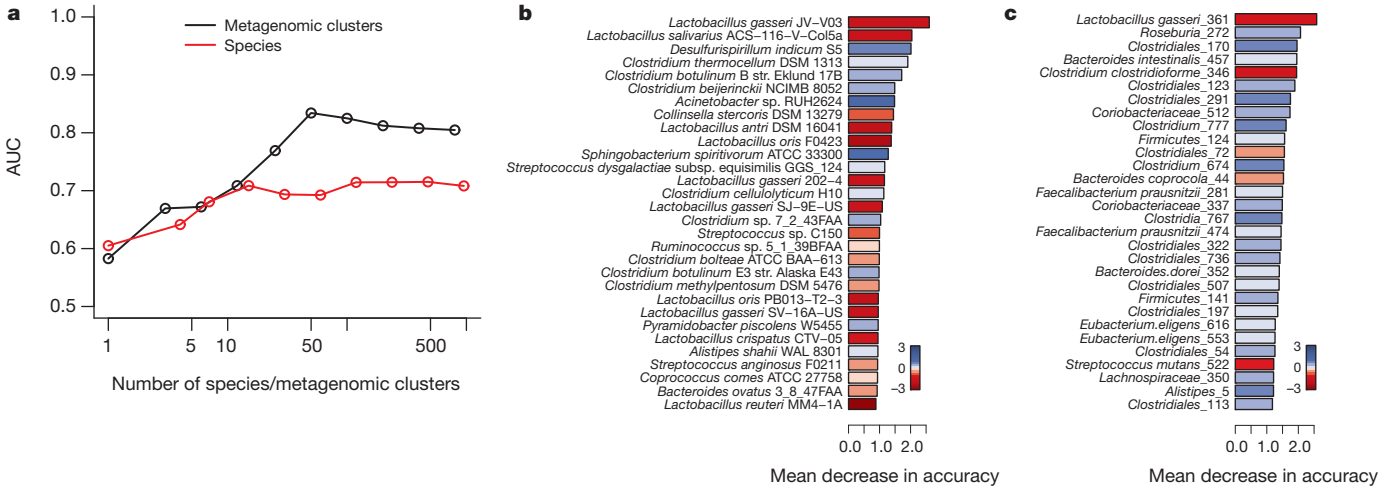


Figure 3 | Classification of T2D status by abundance of species and MGCs. **a**, Classification performance of a random forest model using species or MGC abundance assessed by area under the receiver-operating characteristic curve (AUC). The performance was explored for different numbers of explanatory variables, ordered in importance. **b**, The 30 most discriminant species in the model using 915 species and discriminating between NGT and T2D women. **c**, The 30 most discriminant MGCs in the model using all 800 MGCs and discriminating between NGT and T2D women. The bar lengths in **b** and **c** indicate the importance of the variable, and colours represent enrichment in T2D (red shades) or NGT (blue shades).

which may affect the results. We also tested whether an MGC model trained on one population could classify T2D individuals from the other population. The MGC model based on our cohort had an AUC of 0.58 for the classification of Chinese T2D subjects, whereas the model based on the Chinese cohort had an AUC of 0.66 for the classification of T2D women in our cohort (Supplementary Fig. 16).

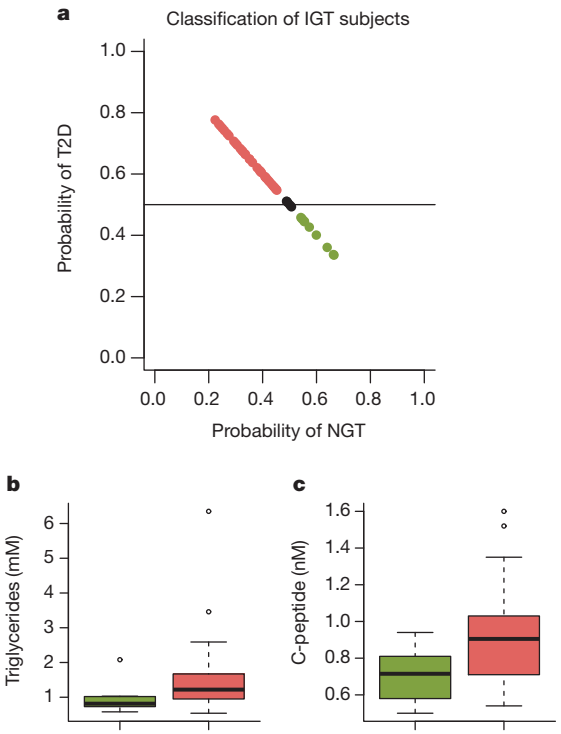


Figure 4 | Stratification of IGT women based on gut microbiota profiles. **a**, Use of the MGC model trained for discriminating NGT and T2D to classify IGT women ($n = 49$) as either NGT (green) or T2D (red). **b**, **c**, IGT women predicted to be T2D had higher triglyceride levels ($P = 0.019$, Wilcoxon rank sum test) (**b**) and higher C-peptide levels ($P = 0.030$, Wilcoxon rank sum test) (**c**). Boxes denote the interquartile range between the first and third quartiles, and the line within denotes the median; whiskers denote the lowest and highest values within 1.5 times interquartile range from the first and third quartiles, respectively. Circles denote data points beyond the whiskers.

These AUC values are lower than the values found both in our work and the previous study[4].

In summary, we characterized the faecal metagenome of 70-year-old European women with T2D, IGT and NGT, and investigated the role of metformin on the microbiome. We also developed the concept of MGCs, which allows DNA that has not previously been sequenced to be included in the analysis. We showed that MGCs identify T2D more accurately than species, indicating that several important gut species still need to be characterized. In addition, we classified women with IGT into subgroups with T2D- or NGT-like metabolism on the basis of their faecal microbiome; this classification offers a potentially new approach to identify individuals at high risk of developing T2D.

Our results are concordant with the recent report showing associations between the gut microbiota and T2D in Chinese individuals[4], despite differences in age. Both studies suggest that functional alterations of the gut microbiome, possibly reflecting changes in the intestinal environment of T2D patients, might be directly linked to T2D development. Although it is likely that the same microbial-encoded functions contribute to disease in different populations, we observed that the most discriminatory MGCs differed between our European T2D subgroup and the Chinese T2D cohort. This observation underscores the need to sample human populations and perform parallel studies in different continents. It also indicates that the development of T2D metagenomic predictive tools and diagnostic biomarkers should be specific to the populations studied.

METHODS SUMMARY

Sample collection. 70-year-old women were recruited using a protocol approved by the ethics committee at Sahlgrenska University Hospital, and were included if they had T2D, IGT or NGT. Exclusion criteria were chronic inflammatory disease and treatment with antibiotics during the preceding 3 months. The subjects were given material and written instructions for providing faecal samples at home.

DNA extraction and sequencing. Genomic DNA was extracted using standard methods[13] and sequenced on Illumina HiSeq 2000. Libraries for each sample were prepared with a fragment length of approximately 300 bp. Low-quality reads and reads mapping to human DNA were removed from the raw data.

Full Methods and any associated references are available in the online version of the paper.

Received 13 December 2012; accepted 17 April 2013.
Published online 29 May 2013.

1. Noble, D., Mathur, R., Dent, T., Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review. *BMJ* **343,** d7163 (2011).

2. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–1031 (2006).
3. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
4. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490,** 55–60 (2012).
5. Karlsson, F. H. *et al.* Symptomatic atherosclerosis is associated with an altered gut metagenome. *Nature Commun.* **3,** 1245 (2012).
6. Mueller, S. *et al.* Differences in fecal microbiota in different European study populations in relation to age, gender, and country: a cross-sectional study. *Appl. Environ. Microbiol.* **72,** 1027–1033 (2006).
7. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* **5,** e10667 (2010).
8. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl Acad. Sci. USA* **108,** 4586–4591 (2011).
9. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
10. De Filippo, C. *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl Acad. Sci. USA* **107,** 14691–14696 (2010).
11. Brohall, G., Behre, C. J., Hulthe, J., Wikstrand, J. & Fagerberg, B. Prevalence of diabetes and impaired glucose tolerance in 64-year-old Swedish women: experiences of using repeated oral glucose tolerance tests. *Diabetes Care* **29,** 363–367 (2006).
12. Fagerberg, B., Kellis, D., Bergstrom, G. & Behre, C. J. Adiponectin in relation to insulin sensitivity and insulin secretion in the development of type 2 diabetes: a prospective study in 64-year-old women. *J. Intern. Med.* **269,** 636–643 (2011).
13. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–134 (2010).
14. Gaetti-Jardim, E. Jr, Marcelino, S. L., Feitosa, A. C., Romito, G. A. & Avila-Campos, M. J. Quantitative detection of periodontopathic bacteria in atherosclerotic plaques from coronary arteries. *J. Med. Microbiol.* **58,** 1568–1575 (2009).
15. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
16. Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C. & Hu, F. B. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men. *Am. J. Clin. Nutr.* **81,** 555–563 (2005).
17. Louis, P., Young, P., Holtrop, G. & Flint, H. J. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. *Environ. Microbiol.* **12,** 304–314 (2010).
18. Vrieze, A. *et al.* Transfer of intestinal microbiota from lean donors increases insulin sensitivity in subjects with metabolic syndrome. *Gastroenterology* **143,** 913–916 (2012).
19. Furet, J. P. *et al.* Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers. *Diabetes* **59,** 3049–3057 (2010).
20. Lundberg, V., Stegmayr, B., Asplund, K., Eliasson, M. & Huhtasaari, F. Diabetes as a risk factor for myocardial infarction: population and gender perspectives. *J. Intern. Med.* **241,** 485–492 (1997).
21. Vendrame, F. & Gottlieb, P. A. Prediabetes: prediction and prevention trials. *Endocrinol. Metab. Clin. North Am.* **33,** 75–92 (2004).
22. Oliveira, A. P., Patil, K. R. & Nielsen, J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.* **2,** 17 (2008).
23. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA* **102,** 2685–2689 (2005).
24. Finegold, S. M. *et al.* Clostridium clostridioforme: a mixture of three clinically important species. *Eur. J. Clin. Microbiol. Infect. Dis.* **24,** 319–324 (2005).
25. Larsen, N. *et al.* Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PLoS ONE* **5,** e9085 (2010).
26. Karjalainen, K. M., Knuuttila, M. L. & Kaar, M. L. Salivary factors in children and adolescents with insulin-dependent diabetes mellitus. *Pediatr. Dent.* **18,** 306–311 (1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to acknowledge R. Perkins for editing the manuscript and M.-L. Ekholm, B. Jannemark, G. Östergren-Lundén, H. Kling Bäckhed, C. Schmidt and U. Prahl for technical assistance. This study was supported by the Swedish Research Council, the NovoNordisk foundation, Torsten Söderberg's foundation, Ragnar Söderberg's foundation, Swedish Diabetes foundation, Swedish Heart Lung Foundation, IngaBritt och Arne Lundbergs foundation, Chalmers foundation, Bioinformatics Infrastructure for Life Sciences (BILS), Knut and Alice Wallenberg foundation, the Swedish Foundation for Strategic Research, AstraZeneca R&D Mölndal, Sweden and the regional agreement on medical training and clinical research (ALF) between Region Västra Götaland and Sahlgrenska University Hospital. The computations were performed on Chalmers C3SE computing resources.

Author Contributions C.J.B., B.F. and F.B. conceived and designed the project. F.H.K., V.T., C.J.B. and G.B. performed the experiments. F.H.K., I.N. and J.N. analysed the sequence data. All authors contributed to writing and editing the manuscript.

Author Information Gut metagenome sequences have been deposited in the Sequence Read Archive (sra) under accession code ERP002469. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.N. (nielsenj@chalmers.se) or F.B. (fredrik.backhed@wlab.gu.se).

METHODS

Study design and recruitment. During 2001–2003, all 64-year-old women in Gothenburg, Sweden, were invited to take part in a screening examination[11] that included anthropometric measurements and a 75-g standardized oral glucose tolerance test, which was repeated in those without NGT. The World Health Organization (WHO) criteria[27] were used for the definitions of diabetes mellitus, IGT and NGT. In the screened cohort of 2,595 women, 9.5% had diabetes mellitus and 14.4% had IGT[11]. As described previously in a study of development of diabetes[12], similarly sized randomized groups with diabetes, IGT and NGT underwent a more detailed baseline examination with a re-examination after more than 5 years. T2D was defined as glutamic acid decarboxylase antibodies < 4.6 units[28].

Women were included in the present sub-study if they had T2D, IGT or NGT. Exclusion criteria were chronic inflammatory disease and treatment with antibiotics during the preceding 3 months. The re-examination took place in 2009 and included questionnaires about current and previous diseases, current medication and smoking habits. Anthropometric measurements were made and blood pressure was recorded. The subjects had fasted overnight and venous blood samples were obtained for measurement of cardiovascular risk factors. We also collected information on medication and glucose control, as well as extensive biometric and plasma measurements. The diagnosis of T2D and IGT at this re-examination was also based on WHO recommendations[27]. All subjects received both written and oral information before they gave their consent to participate in the study. The protocol was approved by the ethics committee at Sahlgrenska University Hospital. After recruitment in the study, three subjects were excluded as they had increased levels of glutamic acid decarboxylase antibodies, indicating type 1 diabetes, and one subject could not be included owing to technical problems with the sequencing.

The characteristics of the included subjects are shown in Supplementary Table 1. The change in glucose tolerance status after a mean of 5.6 years of follow-up is shown in Supplementary Table 2. Supplementary Table 3 lists the biometric and plasma measurements, the country of birth, and the number of years lived in Sweden at the time of first examination for each woman.

The subjects were given material and written instructions for providing faecal samples at home. The samples were produced and transferred to the laboratory one day after the examination. Samples were stored at $-80\,^{\circ}C$ until extraction. Methods for processing faecal samples and isolation of metagenomic DNA have been described previously[13]. DNA concentration was measured with a Nanodrop instrument (Thermo Scientific) and quality was assessed by agarose gel electrophoresis.

Sequencing. All samples were sequenced in the Illumina HiSeq2000 instrument at GATC Biotech with up to 10 samples pooled in one lane. Libraries were prepared with a fragment length of approximately 300 bp. Paired-end reads were generated with 100 bp in the forward and reverse directions.

Data quality control. The length of each read was trimmed with FASTX from the 3′ end of the read using a quality threshold of 20. Read pairs with either reads shorter than 35 bp were removed. Reads that aligned to the human genome (NCBI version 37) (alignment with Bowtie[29], using -n 2 -l 35 -e 200 -best -p 8 -chunkmbs 1024 -X 600 -tryhard) were also removed. This set of high-quality reads was then used for further analysis.

Alignment to reference genomes and taxonomical analysis. The 2,382 microbial reference genomes were obtained from the National Center for Biological Information (NCBI) and Human Microbiome Project (HMP) on 2 August 2011 and were combined into two Bowtie indexes. The metagenomic sequence reads were aligned to reference genomes using Bowtie[29] with the following parameters: -n 2 -l 35 -e 300 -best -p 8 -chunkmbs 1024 -X 600 -tryhard. Mapping results from the two indexes were merged by selecting the alignment with fewest mismatches and a minimum of 90% identities; if a read was aligned to a reference genome with the same number of mismatches, each genome was assigned 1/2 read. The relative abundance of each genome was calculated by summing the number of reads aligned to that genome divided by the total number of reads and scaled by the genome size. In each subject, the relative abundance was scaled to sum to one. The taxonomic rank for every genome (species, genus and phyla) was downloaded from NCBI taxonomy. The relative abundance of taxonomical ranks was calculated by summing the relative abundance of all its members.

***De novo* assembly and gene calling.** High-quality reads were used for *de novo* assembly with Velvet[30] into contigs of at least 500-bp length using a *k*-mer length of 39 coverage cut-off of 3. *k*-mer length was tuned to maximize the N50 value. Reads from each subject were assembled separately; unassembled reads were then used in a global final assembly to also identify rare genes. Genes were predicted on the contigs with MetaGeneMark[31]. A non-redundant gene catalogue was constructed with CD-HIT[32] using a sequence indentity cut-off of 0.95, with a minimum coverage cut-off of 0.9 for the shorter sequences. This catalogue contained 5,997,383 microbial genes (Supplementary Table 20) and was merged with the MetaHIT gene catalogue[15] by adding genes that are unique to our study; the combined gene catalogue was used to align reads. To assess the abundance of genes, reads were aligned to the gene catalogue with Bowtie[29] using parameters: -n 2 -l 35 -e 300 -best -p 8 -chunkmbs 1024 -X 600 -tryhard.

There were 4,778,619 genes unique to our catalogue, which could depend on the fact that we used CD-HIT for clustering whereas BLAT was used in the MetaHIT study[15], although the same criteria of 95% sequence identity and 90% coverage on the shorter sequence were used. Alternatively, this could be owing to the younger age of the MetaHIT cohort (52 ± 11 years (mean $\pm$ s.d.) versus 70 ± 1 years for our cohort; $P < 0.001$, Student's *t*-test), as it is known that the faecal microbiota of adults >65 years is different from that of younger adults[6–8]. We also tested the hypothesis that the increased number of genes in our catalogue could be due to chimaeric or misassembled reads, and tested this hypothesis by applying the method for assembly validation and quality control described previously[33]. This analysis showed that the high number of genes with limited overlap to the MetaHIT catalogue is not due to chimaeric or misassembled reads.

Metagenomic clusters. Genes were clustered based on their profile across samples with the assumption that genes from the same genome should have a similar abundance in one subject. Clustering was done by calculating the correlation distance (1 − correlation coefficient) and clustering with the Markov cluster algorithm implemented in the MCL software[34]. We considered only genes that are shared among 10 or more subjects and calculated the correlation coefficient across subjects, creating edges between genes with values above 0.85. The network was divided into clusters by the MCL software using the suggested values for inflation parameters of 1.4, 2 and 6. Clustering was marginally affected by this change, which suggests a robust clustering. Inflation of 1.4 was chosen for further analysis. Cluster abundance was calculated by summing the relative abundance of all genes in a cluster. To validate clustering, clusters were taxonomically annotated by blasting each gene in a cluster to NCBI non-redundant database with blastp using 1×10^{-5} as *E*-value cut-off. MGCs taxonomical origin (LCA) was determined by blasting the genes in each cluster against the NCBI non-redundant catalogue and requiring that at least 50% of the genes had a best hit to the same phylogenetic group.

Functional annotation. All genes in our catalogue were translated to amino acid sequences and aligned to the KEGG database version 59 using USEARCH[10] ($E < 1 \times 10^{-5}$). Each protein was assigned a KEGG orthologue based on the best hit gene in the KEGG database. Using this approach, 30% of the genes could be assigned a KEGG orthologue and 5,971 unique KEGG orthologues were found. The abundance of a KEGG orthologue was calculated by summing the abundance of genes annotated to a feature.

Statistical analysis. All statistical analyses were made in the R software[35]. Differential abundance of species and MGC was tested by Wilcoxon rank sum test. Correlations between serum biomarkers and species or MGCs were tested with Spearman's correlation. When multiple hypotheses were considered simultaneously, *P* values were adjusted to control the false discovery rate with the method described previously[36]. Only species with a maximum relative abundance in any subject above 10^{-5} was considered in the analyses.

The random forest model has been shown to be a suitable model for exploiting non-normal and dependent data such as metagenomic data[37]. Random forest models were trained using the random forest package in R to identify T2D status in a test set of the NGT and T2D subjects and using the profiles of species and MGCs. The performance of the predictive model was evaluated with a tenfold cross-validation approach and measured as cross-validation error and AUC. Variable importance by mean decrease in accuracy was calculated for the random forest models using the full set of species or MGCs. By ranking the variables by importance, smaller models were constructed containing only the most important variables. The random forest model trained on NGT and T2D subjects was used to classify IGT subjects as NGT or T2D using the profiles of species and MGCs in the random forest package in R with default parameters and 10,000 trees.

For the functional analysis using KEGG orthologues, Wilcoxon rank sum test was used to test differential abundance between groups, and *P* values were corrected for multiple testing with the method described previously[36]. The KEGG grouping of orthologues into pathways was used as input to the reporter feature algorithm[22] and calculating reporter pathways in which there are differential abundant KEGG orthologues. This algorithm takes as inputs the adjusted *P* values and fold changes for each KEGG orthologue in a comparison together with the annotation of KEGG orthologues into pathways from the KEGG database. Each pathway is then scored based on the contributing *P* values of KEGG orthologues and direction by fold changes to calculate a global *P* value for each pathway.

27. Alberti, K. G. & Zimmet, P. Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabet. Med.* **15,** 539–553 (1998).
28. Bingley, P. J., Bonifacio, E. & Mueller, P. W. Diabetes Antibody Standardization Program: first assay proficiency evaluation. *Diabetes* **52,** 1128–1136 (2003).

29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
30. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
31. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38,** e132 (2010).
32. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).
33. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331,** 463–467 (2011).
34. Dongen, S. v. *Graph Clustering by Flow Simulation*. PhD thesis, Univ. Utrecht (2000).
35. R Development Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org (R Foundation for Statistical Computing, 2012).
36. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate — a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* **57,** 289–300 (1995).
37. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35,** 343–359 (2011).

LETTER

doi:10.1038/nature12173

Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants

Dina Lipkind[1,2], Gary F. Marcus[3], Douglas K. Bemis[3], Kazutoshi Sasahara[4], Nori Jacoby[5,6], Miki Takahasi[4], Kenta Suzuki[4,7], Olga Feher[1,2,8], Primoz Ravbar[2,8], Kazuo Okanoya[4,7] & Ofer Tchernichovski[1,2,8]

Human language, as well as birdsong, relies on the ability to arrange vocal elements in new sequences. However, little is known about the ontogenetic origin of this capacity. Here we track the development of vocal combinatorial capacity in three species of vocal learners, combining an experimental approach in zebra finches (*Taeniopygia guttata*) with an analysis of natural development of vocal transitions in Bengalese finches (*Lonchura striata domestica*) and pre-lingual human infants. We find a common, stepwise pattern of acquiring vocal transitions across species. In our first study, juvenile zebra finches were trained to perform one song and then the training target was altered, prompting the birds to swap syllable order, or insert a new syllable into a string. All birds solved these permutation tasks in a series of steps, gradually approximating the target sequence by acquiring new pairwise syllable transitions, sometimes too slowly to accomplish the task fully. Similarly, in the more complex songs of Bengalese finches, branching points and bidirectional transitions in song syntax were acquired in a stepwise fashion, starting from a more restrictive set of vocal transitions. The babbling of pre-lingual human infants showed a similar pattern: instead of a single developmental shift from reduplicated to variegated babbling (that is, from repetitive to diverse sequences), we observed multiple shifts, where each new syllable type slowly acquired a diversity of pairwise transitions, asynchronously over development. Collectively, these results point to a common generative process that is conserved across species, suggesting that the long-noted gap between perceptual versus motor combinatorial capabilities in human infants[1] may arise partly from the challenges in constructing new pairwise vocal transitions.

In the three species we studied, vocal behaviour spans a broad range of combinatorial capabilities: zebra finches sing mostly linear sequences of syllables; Bengalese finch song includes branching sequences; pre-lingual human infants develop a capacity to transition between many syllables, eventually allowing flexible imitation of a potentially infinite array of words[2]. In zebra finches, we tested how the birds solve two combinatorial tasks: swapping syllable order, and inserting syllables into strings. In Bengalese finches, we examined the ontogenetic origin of combinatorial plasticity in specific vocal transitions. In human infants, we examined, statistically, how diversification of many vocal transitions comes about. Across these levels of analysis we tested whether the capacity to rearrange vocal units flexibly is the starting point of vocal learning. Alternatively, the combinatorial machinery might develop slowly, through growth or learning, with individual vocal transitions introduced gradually. Such an early process could enable selective pruning later on. In the first case, we would observe simultaneous and parallel appearance of many syllable transitions during learning; in the latter case, we would observe a stepwise addition of particular transitions to the vocal repertoire.

We trained young zebra finches to imitate playbacks of one song (source), selected birds (17 out of 87) who imitated it fast enough (by day 63 after hatching), and switched their training to a variant song (target)[3,4] where syllable order was altered: ABC–ABC → ACB–ACB (Fig. 1a). We then examined the entire time course of the shift from source to target song (Supplementary Information section 1 and Supplementary Fig. 1)[5,6]. A bigram Markov model was found to account for the bulk of song sequence structure during the experimental period (Supplementary Information section 2 and Supplementary Fig. 2).

In the birds that completed the task ($n = 8$; Fig. 1b–g), the target song appeared abruptly after 17 ± 4.4 days (mean $\pm$ s.e.m., and hereafter; Fig. 1c and Supplementary Fig. 3a). Extinction of the source song occurred before, or concurrently with, the appearance of the target, with a time lag of 3 ± 2.8 days between source disappearance and target first appearance, indicating that the target song was generated by intermediate steps, but with no persistence of old singing habits once the entire target song was in place. To quantify intermediate steps, we tracked the appearance of the target pairwise transitions (bigrams AC, CB and BA), the increase (adjustment) of their frequencies and the extinction of source transitions that were no longer required (Fig. 1d–g).

New transitions appeared sparsely over development, with a lag of 10.4 ± 1.91 days from training onset, and a gap of 6.4 ± 3.5 days between consecutive bigram appearances (Fig. 1d, e, Supplementary Information section 3.1 and Supplementary Fig. 3b, d, e). Each gap included several thousand renditions of a single newly acquired bigram with no concurrent increase in the (zero or near-zero) frequency of target bigrams that were not yet acquired (Supplementary Information sections 3.2–3.3 and Supplementary Fig. 3f–h). Time gaps showed no developmental trend (no significant correlation between first–second and second–third transition gaps; $r^2 = 0.073$). In contrast, both adjustments and extinctions of transitions showed strong developmental trends (Fig. 1g): the appearance of each new target bigram was followed by a fast adjustment to end-point frequency (phase transition), the speed of which increased strongly with the order of bigram appearance: the time interval from 25% to 75% of the end-point frequency was 9.6 ± 2 days for the first bigram, 4.0 ± 0.9 days for the second and only 2.3 ± 0.2 days for the third and final bigram (Fig. 1g, left; $P = 0.018$, paired t-test first versus third bigram). Extinction of source bigrams lagged behind the appearances of first and second target bigrams (4.5 ± 3 and 6 ± 3 days, respectively), but occurred almost simultaneously with the appearance of the third target bigram (-0.3 ± 0.3 days), resulting in a prompt switch to exclusive target performance (Fig. 1g, right, and Supplementary Fig. 3c). The prompt and rapid changes observed once the last bigram appeared probably mirror capabilities not fully expressed earlier (Supplementary Information section 3.4), suggesting that bigram appearance was a rate-limiting stage; namely, once a bigram was performed at all, or above some very low threshold, its frequency could change rapidly to match the target.

[1]Department of Psychology, Hunter College, City University of New York, New York, New York 10065, USA. [2]Department of Biology, City College, City University of New York, New York, New York 10031, USA. [3]Department of Psychology, New York University, New York, New York 10003, USA. [4]Laboratory for Biolinguistics, RIKEN Brain Science Institute, Wako, Saitama 351-0198, Japan. [5]The Interdisciplinary Center for Neural Computation, Hebrew University, 91904 Jerusalem, Israel. [6]The Department of Music, Bar Ilan University, Ramat Gan 5290002, Israel. [7]JST ERATO Okanoya Emotional Information Project, Wako, Saitama 351-0198, Japan. [8]The City University of New York Graduate Center, New York, New York 10016, USA.

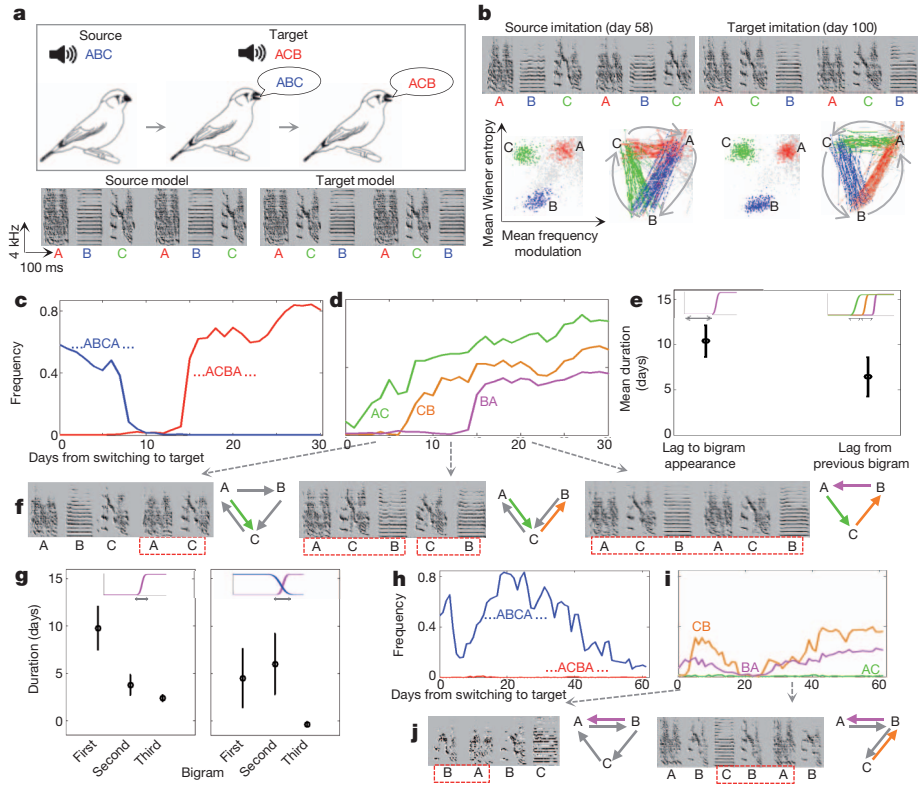


Figure 1 | Syllable rearrangement task. **a**, Top, sequential training with two songs; bottom, training models. **b**, Song examples (top) and scatter plots of syllable features (bottom) after source and after target learning in one bird. Clusters represent syllable types and lines represent transitions (colours represent transition end syllable). **c**, **d**, Daily frequencies (in one bird) of (**c**) source and target songs and (**d**) target bigrams. **e**, Learning phases in successful birds (means ± s.e.m.; $n = 8$). **f**, Songs and syntax diagrams during learning (same bird as in **c**, **d**). **g**, Duration of adjustment (left) and extinction (right) according to bigram appearance order. **h**–**j**, Same as **c**, **d** and **f** in an unsuccessful bird.

From the nine birds that failed to complete the task, five partly learned it (Fig. 1h–j, Supplementary Information section 4 and Supplementary Fig. 4). Their learning process was similar to that of successful birds, except for failing to perform a single (and in one case two) target transition. Consequently, the end point of unsuccessful birds resembled intermediate stages of learning in successful birds (Fig. 1f, j), including a higher transition entropy which merely mirrored the coexistence of source and target bigrams (Supplementary Information section 5 and Supplementary Fig. 5). Thus, despite performing millions of syllable renditions, unsuccessful birds had no measurable capability of producing the entire set of target transitions.

To test if newly acquired syllables can form transitions more easily, we constructed a task that elicited combinatorial changes in newly formed syllables, training birds to incorporate newly formed B syllables into strings of A^n syllables AAAA → ABAB, namely $A^n \rightarrow (AB)^n$ (Fig. 2a, b, Supplementary Information section 6 and Supplementary Fig. 6). Note that syllable B can be inserted into the string even as an unstructured precursor of B. This task was indeed easier: 15 out of 28 birds learned the source song and ten of them also imitated the target song (Fig. 2c, d, Supplementary Information section 6 and Supplementary Fig. 7a). However, birds did not directly insert syllable B into A^n strings; instead, they acquired two new transitions, AB and BA (Fig. 2e and Supplementary Fig. 7b), with an initial delay of 9.7 ± 1.9 days, and time gaps of 4.9 ± 1.52 days between their appearances (Fig. 2h), comparable to appearance gaps in the sequence rearrangement task. As in the ABC → ACB task, adjustment durations tended to decrease with bigram order (7.8 ± 1.6 and 5.1 ± 0.9 days for the first and second bigrams; Fig. 2i), and extinction of the source bigram (AA) usually occurred simultaneously with the appearance of the last target bigram (−1 ± 1.5 days; Fig. 2j and Supplementary Information section 7).

In this task, the newly formed syllable type should initially appear exclusively at the song's edge until it can be 'connected' by two distinct

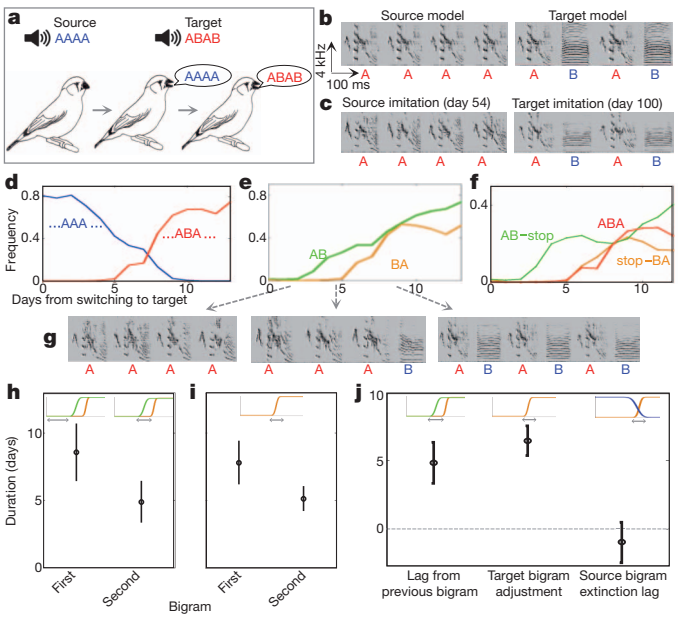


Figure 2 | Syllable insertion task. **a**, **b**, Training regime and models. **c**, Learning outcome in one bird. **d**–**f**, Daily frequencies of syllable sequences in one bird: **d**, source and target songs; **e**, target bigrams; **f**, occurrences of syllable B at bouts' end (green), start (orange) and middle (red). **g**, Song examples during learning (same bird). **h**–**j**, Means ± s.e.m. ($n = 10$) of (**h**) appearance lags of target bigrams, (**i**) adjustment durations and (**j**) lags between target bigrams' appearance, adjustment of the second target bigram and extinction of the source bigram (AA).

bigrams. For example (Fig. 2f, g), if AB is learned first, the bird must stop after singing B, confining B to appear at the end of A^n strings until the second bigram (BA) is learned. To test for such an 'edge effect', we calculated daily frequencies of the occurrence of B at the start of the song (BA^n), at its end (A^nB) and in its middle (ABA). As expected, B was initially performed exclusively at one edge of the song (BA^n in five birds and A^nB in three birds; Fig. 2f, Supplementary Information section 7 and Supplementary Fig. 7c). In all cases, syllable B appeared in the middle of song bouts immediately once the second bigram was learned. Namely, we did not observe cases of a $BA^nB \mid n > 1$ stage before $(AB)^n$, indicating that the only obstacle for incorporating B into the bout centre was inability to perform both AB and BA transitions, as opposed to difficulties in breaking AA transitions. We observed a similar 'edge effect' also in naturally occurring syntax development (Supplementary Information section 8 and Supplementary Fig. 8). Therefore, stepwise acquisition of bigrams generalizes to earlier stages of vocal development, and to a different learning task, where we juxtaposed the formation of a new syllable type with a sequence rearrangement task.

By selecting for fast-learning birds and training them unnaturally, we might have underestimated the full range of combinatorial capabilities that birds might express under more natural conditions. To address this, we studied Bengalese finches, raised in a semi-natural aviary ($n = 8$; Supplementary Information section 9). Although altered-target training was necessary to induce sequence rearrangement in zebra finches, Bengalese finches naturally rearrange syllables as adults[2,7]. We examined the ontogentic origin of song-syntax plasticity by tracing the development both of fixed and of variable parts of the adult song. Consider a case of a bidirectional transition in the mature song A ↔ B (Fig. 3a): this plasticity might be a residual of an early stochastic performance of transitions, including both AB and BA. Alternatively, transitions are acquired sparsely, say AB and later BA. We identified seven bidirectional transitions in the end-point song of five of our experimental birds, and tracked the frequencies of both bigrams (AB and BA) from the earliest time point when both syllable types A and B could be recognized (days 65–83 after hatching) to the end of development (Fig. 3a and Supplementary Fig. 9a). We found long gaps between bigram appearances (17.7 ± 8.7 days; Fig. 3a–c and Supplementary Fig. 9b), and adjustment durations were shorter (8.9 ± 2.2 days; Fig. 3c).

Next, we traced the ontogenetic origin of unidirectional transitions (AB). In 15 out of 16 cases (Supplementary Information section 9), as early as the clusters corresponding to A and B could be identified, significant frequency of AB transitions could be identified, but the frequency of the reverse transitions (BA) was zero or near zero (20 ± 2% versus 1 ± 0.9% for AB and BA, respectively, $P < 0.001$, paired t-test). Therefore, both unidirectional (Fig. 3d, top) and bidirectional (Fig. 3d, bottom) transitions tended to originate from unilateral transitions, which is inconsistent with the notion of highly stochastic transitions early on.

Focusing on bidirectional transitions was necessary to overcome biases in the detection level of syllables during early development: because of symmetry, such biases should not affect the relative frequencies of AB and BA. However, once all syllable types were in place we were able to examine all transition types (Fig. 3e, f). During that period, five out of eight birds kept adding and removing transitions. As in early song development, this process was biased to increase connectivity across syllable types (Fig. 3e; 15 additions versus six deletions across birds) and decrease repetitive sequences (zero additions versus four deletions). Further, looking at branching points, the mean number of variegated transitions (excluding reduplications) per syllable increased over time (3.28 ± 0.24 and 3.88 ± 0.19 for the start and end points, respectively; $P = 0.04$, paired t-test; Fig. 3f). Thus, combinatorial plasticity observed in the adult bird developed from a more restricted syntax, in a stepwise manner.

Finally, we examined the development of phonetic syntax of infant babbling. Classical studies identified a transition from predominantly reduplicated utterances (for example, 'ba ba ba') to variegated utterances (for example, 'ba gu ge')[8,9], which could perhaps mirror a stepwise acquisition of variable transition types. However, later studies failed to replicate this effect reliably[10,11], and instead reported variegated utterances throughout babbling development (Supplementary Information section 10). This could suggest that, unlike songbirds, human infants can rearrange syllables early on with relative ease. However,

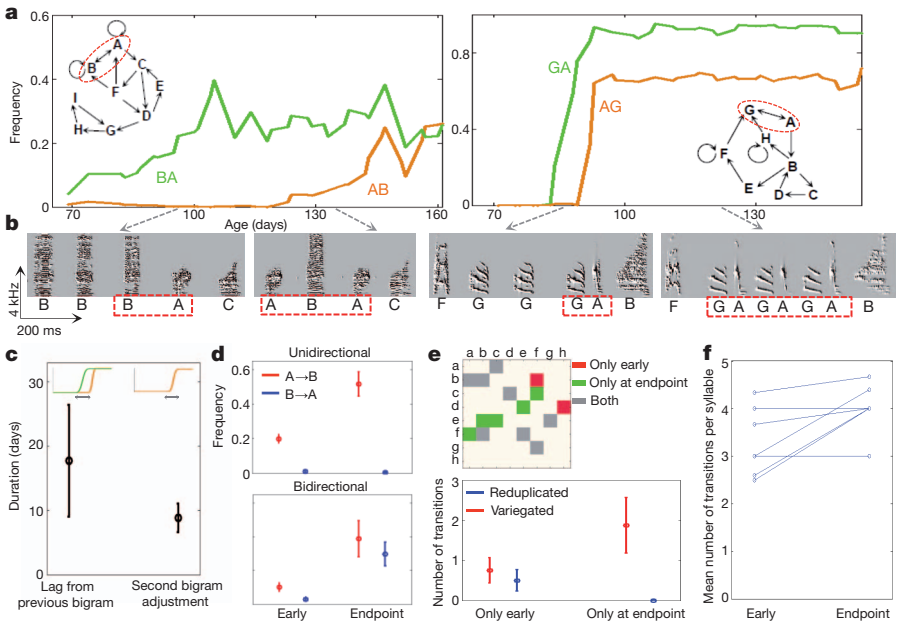


Figure 3 | Combinatorial learning in Bengalese finches. **a**, **b**, Development of bidirectional transitions in two birds. Insets show end-point syntax. **c**, Mean ± s.e.m. of appearance lag and adjustment duration in bidirectional transitions ($n = 7$ transitions). **d**, Means ± s.e.m. of the frequencies of unidirectional (top, $n = 16$) and bidirectional (bottom, $n = 7$) transitions in early development and at end point. **e**, Top, binary transition matrix (one bird), showing transitions present only early (green), only at end point (red) and in both (grey). Bottom, means ± s.e.m. across birds ($n = 8$) of the number of transitions present only early or only at end point. **f**, Developmental changes in the mean number of transitions per syllable (in cases of variable transitions).

infants' large repertoire of syllable types is acquired gradually, so that at any time point the infant is producing a mixture of newly acquired and old syllable types: if for each syllable type the number of available transitions increases gradually, then we would expect less variegation in newly acquired syllables and more in old syllables. The mixture of old and new syllable types at any developmental time could mask a syllable-specific increase in variegation. We therefore tested the existence of a developmental trend in variegation, in reference to the development of specific syllables.

We analysed databases of phonologically transcribed babbling sessions (CHILDES[12,13]) from nine US infants recorded once every 2 weeks at ages 9–28 months, which we segmented into syllables (Methods, see section on 'Analysis of babbling data', and Supplementary Information sections 11.1 and 11.2). We pooled all measures across syllable types in each child, and adjusted our measures through a bootstrapped normalization in each session to control for effects due to developmental changes in the number of syllable types and in utterance length (Supplementary Information section 11.3).

We first tracked the frequencies of reduplicated transitions over infants' ages, aligning the data in reference to the age where the speech/babbling ratio reached 50%. Throughout development, reduplicated utterances were performed $15 \pm 5.7\%$ above chance ($P < 0.001$). However, we did not observe any changes in the tendency to reduplicate syllables over development (Fig. 4a, adjusted $r^2 = 0.01$; $P = 0.32$).

Next, we calculated the same measure again, aligning the data in reference to the appearance time of each syllable type (Fig. 4b). Strikingly, a clear shift from high to low reduplication frequency was now observed (adjusted $r^2 = 0.26$; $P < 0.001$), occurring very slowly, over 20–30 weeks from the time of appearance. Namely, syllables tended to be repeated (reduplicated) when first acquired, and this was followed by a gradual acquisition of transitions to other syllables (variegation). Therefore, previous failures to find a developmental shift from reduplicated to variegated babbling[10,11] may be explained by a masking effect due to asynchronous appearance of new syllable types.

Our findings in songbirds predict that, in infants, new syllable types should first appear at an utterance edge. Indeed, newly generated syllable types appeared more frequently at utterance edges ($8.6 \pm 4\%$ above chance, $P < 0.001$), and this tendency decreased slowly over 20 weeks or so (adjusted $r^2 = 0.13$; $P < 0.01$; Fig. 4c). Finally, we found that the rate of acquiring transitions was lower than expected by chance, taking into account that new syllable types are continuously added to the child's vocabulary, necessarily enlarging the pool of potential syllable transitions ($28 \pm 8\%$ below chance at first session; $P < 0.001$; Fig. 4d). Namely, the increase in the number of bigrams lagged behind the increase in syllable vocabulary.

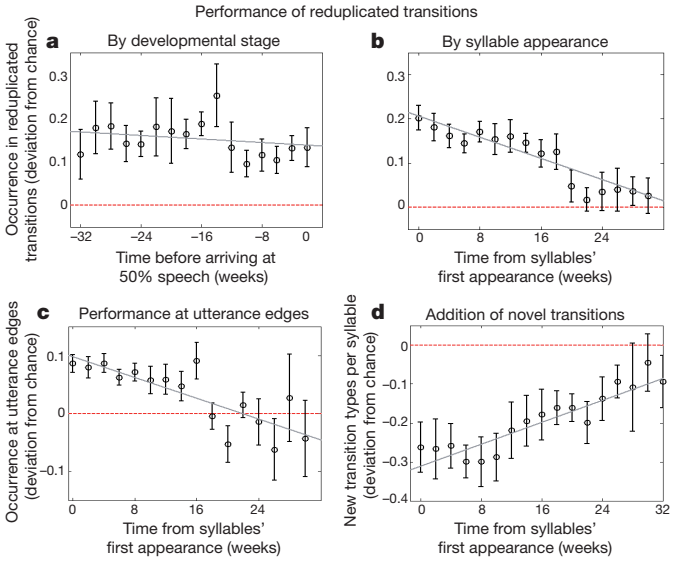


Figure 4 | Incorporation of new syllables into infants' babbling utterances. **a**, **b**, Frequency of syllable occurrence in reduplicated transitions: (**a**) data aligned by developmental stage (time zero is the first session with more than 50% of speech utterances); (**b**) data aligned by each syllable type's first appearance. **c**, Frequency of syllable occurrence at utterance edges. **d**, New transition types added per syllable type. **a**–**d**, Means ± s.e.m. across children ($n = 9$) are deviations from chance level (zero, red dashed line, assessed by bootstrap analysis). Grey lines, fitted linear model.

Our results across species suggest that, in contrast to predictions of previous theories[14,15], new vocal transitions are acquired slowly during early stages of development. A similar, gradual, generative process was observed in the development of non-learned movement sequences[16,17], although intriguingly, not in studies of movement sequence learning in adult monkeys, where target sequences appeared very rapidly but frequency adjustments took weeks[18,19]. Whether these differences are due to age, experience or learning modules (movement versus vocal) should be investigated in future work.

Our findings point to a prolonged developmental stage of stepwise acquisition of vocal combinatorial capacity, which may be accompanied or followed by pruning of unnecessary transitions[20,21]. Dynamics of a trial-and-error learning alone are unlikely to explain the zero or near-zero frequencies of many transitions for prolonged developmental epochs (Supplementary Information section 12). Instead, we propose that stepwise development of combinatorial diversity might stem from the dynamics of constructing new links between representations of vocal gestures in the motor system: In songbirds, vocal production gradually differentiates into distinct syllable types, represented by chains of neuronal activity[22] in the motor song system, which are thought to code sequences of vocal gestures[23]. During singing, neuronal activity must propagate from the tail of one chain of gestures to the head of the other[24]. The construction of such connections might be initially sparse, limiting vocal sequences to a small set of transitions, and reduplications. Adding and removing tail-to-head connections should allow additions and deletions of vocal transitions (AB↔ABC) but not swapping (ABC→ACB) or insertions (AA→ABA). If the process is dominated by additions, we should see more and more branching sequences (as in Bengalese finches), and eventually (perhaps in human infants) an all-to-all network with a single connected component might emerge, allowing free access to any element, which is later pruned to produce speech[20,21]. According to this model, vocal babbling is shaped by a slowly evolving inter-syllabic network, where freedom gained due to acquiring new transitions is counterbalanced by the acquisition of new syllable types that are not yet connected and that tend to reduplicate or break the sequence. Such a process could explain the mismatch between infants' precocious ability to perceive complex grammars, and their initially limited ability to produce vocal sequences[1]. A similar gap may also exist in songbirds, whose perceptual capacity for syntax learning is a debated question[25–29]. However, there is also a fascinating parallel between the perceptual ability of songbirds to assemble memories of phrase pairs into a complete multi-phrase song template[30], and the phenomenon shown here, of birds and infants using pairwise syllable transitions to transform one multi-syllable string into another.

METHODS SUMMARY

Animal care. All experiments were conducted in accordance with the guidelines of the US National Institutes of Health and were reviewed and approved by the Institutional Animal Care and Use Committees of Hunter College and City College of the City University of New York, and of RIKEN Brain Science Institute.
Experimental design. Male zebra finches were reared and housed singly, in sound attenuation chambers as previously described[5]. Audio-recording and training with song playbacks used Sound Analysis Pro[6]. Bengalese finches were reared in communal aviaries in RIKEN Brain Science Institute, Japan. Starting from 40 to 50 days of age, they were transferred singly to soundproof cages at 3- to 4-day intervals and recorded for approximately 24 h.

Data analysis. We used Sound Analysis Pro[6] for song feature calculation and cluster analysis. We used Matlab for the rest of the analysis. Cluster information was used to elucidate the order of syllable types sung. Daily frequencies of sequences (source and target songs and bigrams) were calculated as the proportion of syllables constituting the sequence out of the total number of syllables sung on that day.

Infant babbling data were obtained from the Davis corpus[13] of the CHILDES database[12]. Only babbling utterances were analysed. Utterances were semi-athomatically parsed into syllables (Methods, see section on 'Analysis of babbling data'). Bootstrap normalization was used to establish a value for each measure reflecting a random placement of syllables with vocabulary size and utterance length held constant (Methods, see section on 'Analysis of babbling data'). Measures were calculated for each syllable type in each session, and a mean over syllable types was calculated for each time point.

Full Methods and any associated references are available in the online version of the paper.

Received 5 April 2012; accepted 8 April 2013.
Published online 29 May 2013.

1. Marcus, G. F., Vijayan, S., Bandi Rao, S. & Vishton, P. M. Rule learning by seven-month-old infants. *Science* **283,** 77–80 (1999).
2. Berwick, R. C., Okanoya, K., Beckers, G. J. L. & Bolhuis, J. J. Songs to syntax: the linguistics of birdsong. *Trends Cogn. Sci.* **15,** 113–121 (2011).
3. Eales, L. Song learning in zebra finches: some effects of song model availability on what is learnt and when. *Anim. Behav.* **33,** 1293–1300 (1985).
4. Plamondon, S. L., Rose, G. J. & Goller, F. Roles of syntax information in directing song development in white-crowned sparrows (*Zonotrichia leucophrys*). *J. Comp. Psychol.* **124,** 117–132 (2010).
5. Derégnaucourt, S., Mitra, P. P., Fehér, O., Pytte, C. & Tchernichovski, O. How sleep affects the developmental learning of bird song. *Nature* **433,** 710–716 (2005).
6. Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. & Mitra, P. P. A procedure for an automated measurement of song similarity. *Anim. Behav.* **59,** 1167–1176 (2000).
7. Yamashita, Y. *et al.* Developmental learning of complex syntactical song in the Bengalese finch: a neural network model. *Neural Netw.* **21,** 1224–1231 (2008).
8. Oller, D. K. in *Child Phonology* Vol. 1 (eds Yeni-Komshian, G., J. Kavanagh, J. & Ferguson, C.) 93–112 (Academic, 1980).
9. Stark, R. in *Child Phonology* Vol. 1 (eds Yeni-Komshian, G., J. Kavanagh, J. & Ferguson, C.) 73–92 (Academic, 1980).
10. Mitchell, P. R. & Kent, R. D. Phonetic variation in multisyllable babbling. *J. Child Lang.* **17,** 247–265 (1990).
11. Smith, B. L., Brown-Sweeney, S. & Stoel-Gammon, C. A quantitative analysis of reduplicated and variegated babbling. *First Lang.* **9,** 175–189 (1989).
12. MacWhinney, B. The CHILDES project: tools for analyzing talk. *Child Lang. Teach. Ther.* **8,** 217–218 (1992).
13. Davis, B. L. & MacNeilage, P. F. The articulatory basis of babbling. *J. Speech Hear. Res.* **38,** 1199–1211 (1995).
14. Edelman, G. *Neural Darwinism. The Theory of Neuronal Group Selection* (Basic Books, 1987).
15. Hanuschkin, A., Diesmann, M. & Morrison, A. A reafferent and feed-forward model of song syntax generation in the Bengalese finch. *J. Comput. Neurosci.* **31,** 509–532 (2011).
16. Golani, I. A mobility gradient in the organization of vertebrate movement?: the perception of movement through symbolic language. *Behav. Brain Sci.* **15,** 249–308 (1992).
17. Dominici, N. *et al.* Locomotor primitives in newborn babies and their development. *Science* **334,** 997–999 (2012).
18. Hikosaka, O., Rand, M. K., Miyachi, S. & Miyashita, K. Learning of sequential movements in the monkey: process of learning and retention of memory. *J. Neurophysiol.* **74,** 1652–1661 (1995).
19. Rand, M. K. *et al.* Characteristics of sequential movements during early learning period in monkeys. *Exp. Brain Res.* **131,** 293–304 (2000).
20. De Boysson-Bardies, B. & Vihman, M. M. Adaptation to language: evidence from babbling and first words in four languages. *Language* **67,** 297–319 (1991).
21. Vihman, M. M., Macken, M. A., Miller, R., Simmons, H. & Miller, J. From babbling to speech: a re-assessment of the continuity issue. *Language* **61,** 397–445 (1985).
22. Jin, D. Z., Ramazanoğlu, F. M. & Seung, H. S. Intrinsic bursting enhances the robustness of a neural network model of sequence generation by avian brain area HVC. *J. Comput. Neurosci.* **23,** 283–299 (2007).
23. Amador, A., Perl, Y. S., Mindlin, G. B. & Margoliash, D. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* **495,** 59–64 (2013).
24. Jin, D. Z. Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. *Phys. Rev. E* **80,** 051902 (2009).
25. Abe, K. & Watanabe, D. Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature Neurosci.* **14,** 1067–1074 (2011).
26. Beckers, G. J. L., Bolhuis, J. J., Okanoya, K. & Berwick, R. C. Birdsong neurolinguistics: songbird context-free grammars claim is premature. *NeuroReport* **23,** 139–145 (2012).
27. Gentner, T. Q., Fenn, K. M., Margoliash, D. & Nusbaum, H. C. Recursive syntactic pattern learning by songbirds. *Nature* **440,** 1204–1207 (2006).
28. Katahira, K., Suzuki, K., Okanoya, K. & Okada, M. Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. *PLoS ONE* **6,** e24516 (2011).
29. Van Heijningen, C. A. A., de Visser, J., Zuidema, W. & ten Cate, C. Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. *Proc. Natl Acad. Sci. USA* **106,** 20538–20543 (2009).
30. Rose, G. J. *et al.* Species-typical songs in white-crowned sparrows tutored with only phrase pairs. *Nature* **432,** 753–758 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Benichov, J. Hyland Bruno, I. Ljubičić and C. Roeske for help with data analysis. We also thank A. Vouloumanos, M. Hauber, L. Parra and V. Valian for reading the manuscript. The study was supported by a US Public Health Service grant to O.T., by a Grant in Aid from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan to K.O., and by a Grant in Aid for Japan Society for the Promotion of Science Fellows to M.T.

Author Contributions D.L., O.T., G.F.M., D.K.B., Ka.S. and K.O. designed the research. D.L., O.F. and O.T. performed experiments on zebra finches. D.L., G.F.M., O.F., P.R., N.J. and O.T. analysed data of zebra finches. Ka.S., M.T., Ke.S. and K.O. designed and conducted experiments on Bengalese finches. D.L., Ka.S. and O.T. analysed data of Bengalese finches. D.K.B. analysed infant babbling data, with contributions from G.F.M., O.T. and D.L. D.L., G.F.M., D.K.B., K.O. and O.T. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.L. (dina.lipkind@gmail.com).

METHODS

Animal care. Experiments were conducted in accordance with the guidelines of the US National Institutes of Health and were reviewed and approved by the Institutional Animal Care and Use Committees of Hunter College and City College of the City University of New York, and of RIKEN Brain Science Institute.

Experimental design. Male zebra finches were bred at Hunter College and City College of the City University of New York, and reared in the absence of adult males between days 7 and 30 after hatching. Afterwards, birds were kept singly in sound attenuation chambers and recorded continuously. Twelve out of seventeen birds were passively exposed to 30 playbacks per day of the source, occurring at random with a probability of 0.01 per second, from days 33 to 39 until day 43, in an attempt to increase success rate. On day 43, each bird was trained to press a key to hear song playbacks, with a daily quota of 20. Once birds learned the source, we switched to playbacks of the target. Only birds that learned the source before day 63 ($n = 17$, 20% of the total birds trained with the source) were used. Recording and training used Sound Analysis Pro[6].

Source and target song models were synthetically composed of natural syllables. To balance the design of the sequence rearrangement task, we trained some birds ($n = 7$) with ABC–ABC → ACB–ACB as source and target, and others ($n = 10$) with ACB–ACB → ABC–ABC. The two groups were pooled, and for simplicity we refer to all as ABC–ABC → ACB–ACB.

Bengalese finches were reared in communal aviaries in RIKEN Brain Science Institute, Japan. From the age of 40–50 days, they were transferred singly into a soundproof cage at 3- to 4-day intervals and recorded for approximately 24 h.

Data analysis (songbirds). Song feature calculation and cluster analysis were performed using Sound Analysis Pro[6]. We used MATLAB 7 for further analysis. Cluster information was used to elucidate the order of syllable types sung by a bird on each developmental day and to test whether syllable types were reused in the learning of new syntax (Supplementary Information section 1).

In zebra finches, the percentage of clustered syllables in bouts (assessed by manual inspection of a sample of ten song bouts per bird) was 96 ± 3% at the end point, and 90 ± 2% during the transition from source to target. Clustering Bengalese finch songs was more difficult, with 91 ± 1% at the end point and 80 ± 3% at the starting point. Unidentified syllable types were regarded as missing data.

Song bouts were defined as sequences of identified syllable types with stop durations of less than a threshold that was determined by the typical stop duration in the end-point song (150 ms for ABC–ABC → ACB–ACB and Bengalese finches; 100 ms for AAAA → ABAB). The threshold for bigram occurrences was similarly defined by the typical stop duration in each bigram type at the end point (60–150 ms). Daily frequencies of sequences (source and target songs and bigrams) were calculated as the proportion of syllables constituting a given sequence out of the total number of syllables sung on that day. Because of unavoidable misclassifications in the clustering process, we had to determine a margin of error to decide when an observed transition frequency was real. We empirically estimated our error level as about 2% (2.2 ± 0.5%) by measuring the baseline levels of target bigrams on day 0, and set our threshold to detect the moment of appearance of a bigram transition at 3% above noise (that is, 5%). In an effort to assess the real performance rate of target transitions below noise level, we visually examined a sample of positively identified bigrams on days where their frequency was close to zero (Supplementary Information section 3.2), and found that actual performance rates ranged from very low (0.01) to absolute zero.

Bengalese finches' songs contained more syllable types than those of the zebra finches (6–10 versus 2–3), resulting in a higher level of clustering errors. We therefore took a semi-automatic approach to determine the bigram detection threshold. In the first stage, we used 5%, as in zebra finches; in the second stage, we excluded from our analysis transitions that were clearly an outcome of clustering errors, on the basis of visually examining 20 random instances of each transition type.

Analysis of babbling data. Data were obtained from nine children in the Davis corpus[13] of the CHILDES database[12]. On average, children were 9 months 28.3 days old (s.d. 2 months 1.3 days) at the first session, and data were collected for an average of 1 year 7 months (s.d. 7 months 12.8 days). Data consisted of 38.8 sessions on average per child (s.d. 10.2), recorded an average of 16.07 days apart (s.d. 6.4).

Only babbling utterances (that is, utterances for which no lexical items were assigned in the CHILDES transcriptions) were analysed. Utterances were parsed into syllables using a semi-automated method, described below. Only utterances that received a complete syllabic parse were analysed (2135 utterances per child (s.d. 924) and 62.0 utterances per session (s.d. 37.5)).

We used an iterative parsing process. An utterance was considered parsed if every phoneme in it was successfully assigned to a syllable by the algorithm, such that each phoneme was used exactly one time in a syllable. On each iteration, we first manually assigned complete syllabic parses to several unparsed utterances. We then added new syllable types to the set of possible syllables that could be used for parsing. Next, we automatically checked if all utterances could be exhaustively parsed using the current store of syllables. For example, an utterance 'badaja' would be manually assigned the syllabic parse of 'ba', 'da' and 'ja'. On the following iteration, an utterance 'baja' could be parsed into the syllables 'ba' and 'ja'. Utterances that could not be fully parsed using the set of defined syllables were manually parsed, adding to the set of acceptable syllables. Thus, every syllable used to parse the data was manually verified as a valid syllable in the data. If an utterance could be assigned two different parses, we used a heuristic such that we chose the parse with the greater number of two phoneme syllables (CV or VC). If several parses for an utterance were equal in this measure, we would manually assign a parse to the utterance or leave it as ambiguous, and exclude it from the analysis. Iterations were performed until a sizeable amount of the data had been parsed (58.2% of babbling utterances (s.d. 19.0)).

From this set of parsed utterances, we tabulated the frequency of each syllable in each session and its placement in an utterance. We restricted analysis to syllables that reached a frequency threshold of 1% of the total number of syllables in the session, thus focusing on syllables that the child produced at a non-negligible rate. We also calculated the frequency of all transitions between the syllables. A transition was defined as any two sequential syllables in an utterance. On average, each child used 128 distinct syllables (s.d. 8.12) and constructed 763 distinct transitions (s.d. 95.6).

Measures of the development of transition variability over time are affected by the growth in the number of syllable types and in utterance length. To control for this, we used a bootstrapped normalization procedure for all measures. To establish a baseline value for each measure that reflected a random placement of syllables but held vocabulary size and utterance length constant, we shuffled syllables randomly in each recording session while maintaining the length of each utterance. Each measure was then recalculated over these bootstrapped randomizations to establish a baseline value, to which the observed data were compared.

All measures were calculated for each syllable type in each session. Sessions were then aligned on the first appearance of syllable types, and a mean over syllable types was calculated for each session. For each measure, we evaluated trends over sessions by fitting a simple linear regression model to subject averages using R (R Development Core Team 2007, available at http://cran.r-project.org). Separate models were fitted for each measure with session number as a fixed factor. Only syllable types that appeared in the course of the experimental period (namely, that were not present at the first session) were analysed.

Reduplication was the frequency of occurrences in reduplicated transitions per syllable type. This measure was calculated twice using two different alignments: by developmental stage (the first session where speech/babbling ratio reached 50%) and by the first appearance of the syllable type. Note that the sample size for the developmental alignment analysis was smaller ($n = 7$) than for the syllable-specific alignment ($n = 9$), because of an insufficient number of sessions with more than 50% babbling utterances in two children.

Occurrence of new syllables at edges was the frequency of occurrences of a syllable type at the edge of an utterance compared with occurrences in its middle. For this measure, we did not count reduplications as being in the middle of utterances.

Addition of new transitions was the number of new transition types per syllable type in each session. This measure indicated how likely each syllable occurrence was to participate in a previously unseen transition.

LETTER

doi:10.1038/nature12154

A key role for mitochondrial gatekeeper pyruvate dehydrogenase in oncogene-induced senescence

Joanna Kaplon[1], Liang Zheng[2]*, Katrin Meissl[1]*, Barbara Chaneton[2], Vitaly A. Selivanov[3,4], Gillian Mackay[2], Sjoerd H. van der Burg[5], Elizabeth M. E. Verdegaal[5], Marta Cascante[3,4], Tomer Shlomi[6,7], Eyal Gottlieb[2] & Daniel S. Peeper[1]

In response to tenacious stress signals, such as the unscheduled activation of oncogenes, cells can mobilize tumour suppressor networks to avert the hazard of malignant transformation. A large body of evidence indicates that oncogene-induced senescence (OIS) acts as such a break, withdrawing cells from the proliferative pool almost irreversibly, thus crafting a vital pathophysiological mechanism that protects against cancer[1–5]. Despite the widespread contribution of OIS to the cessation of tumorigenic expansion in animal models and humans, we have only just begun to define the underlying mechanism and identify key players[6]. Although deregulation of metabolism is intimately linked to the proliferative capacity of cells[7–10], and senescent cells are thought to remain metabolically active[11], little has been investigated in detail about the role of cellular metabolism in OIS. Here we show, by metabolic profiling and functional perturbations, that the mitochondrial gatekeeper pyruvate dehydrogenase (PDH) is a crucial mediator of senescence induced by $BRAF^{V600E}$, an oncogene commonly mutated in melanoma and other cancers. $BRAF^{V600E}$-induced senescence was accompanied by simultaneous suppression of the PDH-inhibitory enzyme pyruvate dehydrogenase kinase 1 (PDK1) and induction of the PDH-activating enzyme pyruvate dehydrogenase phosphatase 2 (PDP2). The resulting combined activation of PDH enhanced the use of pyruvate in the tricarboxylic acid cycle, causing increased respiration and redox stress. Abrogation of OIS, a rate-limiting step towards oncogenic transformation, coincided with reversion of these processes. Further supporting a crucial role of PDH in OIS, enforced normalization of either PDK1 or PDP2 expression levels inhibited PDH and abrogated OIS, thereby licensing $BRAF^{V600E}$-driven melanoma development. Finally, depletion of PDK1 eradicated melanoma subpopulations resistant to targeted BRAF inhibition, and caused regression of established melanomas. These results reveal a mechanistic relationship between OIS and a key metabolic signalling axis, which may be exploited therapeutically.

We compared the metabolism of human diploid fibroblasts (HDFs) undergoing OIS to that of cycling cells. To evoke OIS, we used oncogenic $BRAF^{V600E}$, a common oncogene and strong inducer of OIS both *in vitro* and *in vivo*[12–16]. We first analysed the oxygen consumption rate (OCR). Compared to cycling cells, 'OIS cells' showed a significant increase in the OCR, indicating increased mitochondrial oxidative metabolism (Fig. 1a). This was further supported by exchange rate (uptake or secretion) measurements of key metabolites: OIS cells secreted less than half the amount of pyruvate compared with cycling cells, and this was balanced by neither a significant change in glucose consumption nor an increase in lactate or alanine secretion (Fig. 1b, c and Supplementary Fig. 2a, b). Whereas glutamate secretion was increased in OIS cells (Fig. 1d), glutamine consumption was decreased (Supplementary Fig. 2c). In agreement with increased tricarboxylic acid (TCA) cycle activity in OIS cells, stable isotope labelling with a uniformly labelled [U-$^{13}C_6$]glucose tracer showed faster accumulation of glucose-derived 2-carbon-labelled metabolic isotopomers of the TCA cycle (citrate, α-ketoglutarate and malate) as well as TCA cycle-derived 2-carbon-labelled glutamate (Supplementary Fig. 2d). Together, these results indicate that OIS is accompanied by increased pyruvate oxidation, which was corroborated by a simultaneous rise in redox stress, as measured by increased reactive oxygen species (ROS) production, a decrease in the reduced/oxidized glutathione (GSH/GSSG) ratio, and induction of several ROS-responsive genes (Supplementary Fig. 3a–i).

The gatekeeping enzyme linking glycolysis to the TCA cycle is PDH[17,18], which is regulated by reversible phosphorylation[19–22]: phosphorylation by PDK enzymes (PDK1–4) inhibits its action and halts pyruvate use in the TCA cycle, whereas dephosphorylation by PDP1 and PDP2 stimulates PDH activity. Our metabolic profiles predicted PDH to be activated in OIS. Furthermore, if increased PDH activity has a causal role in mediating OIS, we expected this to be reversed after

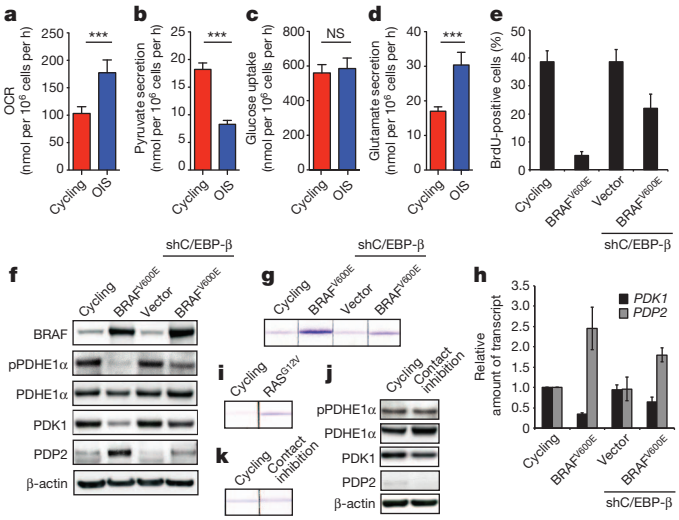


Figure 1 | The PDK1–PDP2–PDH axis is deregulated in OIS. a–d, Analysis of the OCR (**a**), pyruvate secretion (**b**), glucose uptake (**c**) and glutamate secretion (**d**) in cycling and OIS HDFs (after expression of $BRAF^{V600E}$); $n = 6$. **e–h**, Cycling cells and cells undergoing OIS or abrogating OIS (using shC/EBP-β) were analysed for BrdU incorporation (**e**), by immunoblotting (**f**), for PDH activity (**g**) and for regulation of *PDK1* and *PDP2* transcripts, as determined by quantitative reverse transcriptase PCR (qRT–PCR) (**h**); $n = 3$. pPDHE1α denotes phosphorylated PDHE1α. **i**, PDH activity of HDFs undergoing RAS^{G12V}-induced senescence. **j**, **k**, Immunoblotting analysis (**j**) or PDH activity (**k**) of cycling and quiescent HDFs. All data represent mean ± s.d. $^{***}P < 0.001$.

[1]Division of Molecular Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [2]Cancer Research UK, Beatson Institute for Cancer Research, Switchback Road, Glasgow G61 1BD, UK. [3]Department of Biochemistry and Molecular Biology and IBUB, Faculty of Biology, Universitat de Barcelona, Av Diagonal 643, 08028 Barcelona, Spain. [4]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Rosselló 149-153, 08036 Barcelona, Spain. [5]Experimental Cancer Immunology and Therapy, Department of Clinical Oncology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands. [6]Computer Science Department, Technion, Israel Institute of Technology, Haifa 32000, Israel. [7]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA.

*These authors contributed equally to this work.

OIS escape (by C/EBP-β depletion[13] or SV40 small-t-antigen expression; Fig. 1e and Supplementary Fig. 4a). Indeed, PDH phosphorylation (on Ser 293, one of three phosphorylation sites inactivating PDH) was strongly reduced in OIS cells, but this was restored after senescence abrogation (Fig. 1f and Supplementary Fig. 4b). These changes in PDH phosphorylation translated into corresponding alterations in its activity (Fig. 1g).

Notably, in OIS cells, the two key PDH-modifying enzymes were regulated in opposite directions: whereas PDK1 was downregulated, PDP2 was induced (Fig. 1f, h). After OIS abrogation, their levels were normalized. Other PDK and PDP isozymes were not regulated in this manner (Supplementary Fig. 4c, d). Similar to cells undergoing $BRAF^{V600E}$-induced senescence, RAS^{G12V}-senescent cells showed increased accumulation of glucose-derived TCA cycle metabolites (Supplementary Fig. 5a), induction of the OCR (Supplementary Fig. 5b), a drop in PDH phosphorylation (albeit moderately, Supplementary Fig. 5c) and increased PDH enzymatic activity (Fig. 1i), but this was not accompanied by effects on PDK1 or PDP2 levels (Supplementary Fig. 5c). In quiescent cells, labelling of pyruvate, lactate, alanine and citrate from glucose-derived carbons was synchronously increased over time at a higher rate than in proliferating cells (Supplementary Fig. 6a, b), as previously reported[23]. However, the alterations in PDH activity were specific for OIS rather than a consequence of cell cycle arrest, as neither PDH phosphorylation nor PDH enzymatic activity was altered in quiescent cells (Fig. 1j, k). These results indicate that OIS, and escape thereof, is accompanied by antagonistic regulation of two key enzymes controlling PDH activity, PDP2 and PDK1.

To determine whether deregulation of the PDK1–PDP2–PDH axis drives OIS-associated metabolic rewiring, we depleted PDP2 (the expression of which is induced by $BRAF^{V600E}$). This reversed the decrease in PDH phosphorylation in OIS, leading to suppressed PDH activity (Fig. 2a, b). Consistently, [U-$^{13}C_6$]glucose and [$^{13}C_3$]pyruvate labelling revealed that PDP2-depleted cells had less labelling of citrate (Supplementary Figs 7a and 8). In agreement, the ratio of labelled [$^{13}C_2$]citrate (emulating PDH product) to [$^{13}C_3$]pyruvate (PDH substrate), indicating intracellular PDH activity, was reversed by PDP2 depletion (Fig. 2c). These cells also showed decreased OCR and redox stress compared with OIS cells (Fig. 2d and Supplementary Fig. 3b–i).

We next investigated whether PDP2 regulates OIS. Indeed, its depletion abrogated $BRAF^{V600E}$-induced arrest, which was not explained by loss of $BRAF^{V600E}$ signalling (Fig. 2a, e, f). Senescence bypass was accompanied by a marked reduction in the levels of the senescence-associated tumour suppressor $p15^{INK4B}$ (Fig. 2a) and decreased activity of the senescence-associated β-galactosidase (SA-β-gal) biomarker (Fig. 2g). Furthermore, the induction of the OIS-associated and C/EBP-β-dependent interleukins (IL)-6 and 8 (ref. 13) was curtailed by PDP2 depletion (Fig. 2h).

Because PDK1 (the expression of which is suppressed by $BRAF^{V600E}$) antagonizes PDP2 in regulating PDH activity, we also addressed its role in OIS. Ectopic restoration of PDK1 rescued the decrease in PDH phosphorylation and suppressed the increase in PDH activity in OIS cells (Fig. 3a, b). Consistently, PDK1 expression reversed the increase in TCA cycle activity in OIS cells and blocked the rise in PDH activity as judged from the [$^{13}C_2$]citrate:[$^{13}C_3$]pyruvate ratio (Fig. 3c and Supplementary Figs 7b and 8). These effects were mirrored by significant decreases in the OCR and redox stress (Fig. 3d and Supplementary Fig. 3b–i). Restoration of PDK1 expression abrogated the induction of cell cycle arrest after $BRAF^{V600E}$ expression (Fig. 3e, f), which was paralleled by suppression of several senescence-associated biomarkers (Fig. 3a, g, h). The OIS-associated metabolic rewiring was specific to the PDK1–PDP2–PDH axis: depletion of lactate dehydrogenase A

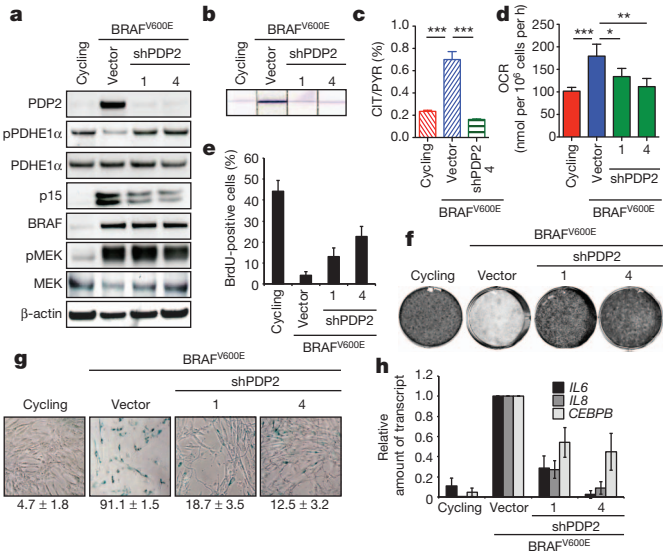


Figure 2 | PDP2 regulates metabolic rewiring and OIS. **a**, Immunoblotting analysis of HDFs expressing empty vector or shPDP2 in the presence or absence of $BRAF^{V600E}$. **b**–**g**, Cells from **a** were analysed for PDH activity in cell extracts (**b**) or intracellularly, as denoted by the [$^{13}C_2$]citrate/[$^{13}C_3$]pyruvate (CIT/PYR) ratio (**c**), the OCR (**d**), BrdU incorporation (**e**), cell proliferation (**f**) and SA-β-gal activity (**g**); $n = 3$. **h**, Regulation of *IL8*, *IL6* and *CEBPB* (also known as *C/EBP-β*) transcripts of the samples described in **a**, as determined by qRT–PCR; $n = 3$. All data are represented as mean ± s.d. $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

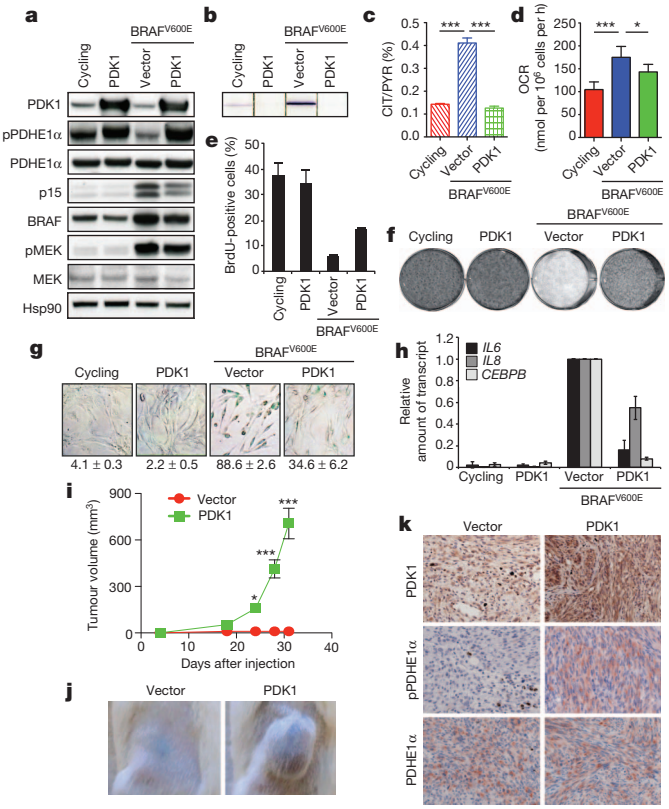


Figure 3 | PDK1 regulates metabolic rewiring and OIS, and acts tumorigenically. **a**, Immunoblotting analysis of HDFs expressing empty vector or PDK1 in the presence or absence of $BRAF^{V600E}$. **b**–**h**, Cells from **a** were analysed for PDH activity in cell extracts (**b**), or intracellularly, as denoted by the [$^{13}C_2$]citrate/[$^{13}C_3$]pyruvate ratio (**c**), the OCR (**d**), BrdU incorporation (**e**), cell proliferation (**f**), SA-β-gal activity (**g**) and *IL8*, *IL6* and *CEBPB* transcript levels (**h**); $n = 3$. **i**, Growth curve of tumours formed by p53-depleted $Braf^{V600E}$-expressing melanocytes and expressing either empty vector or PDK1; $n = 8$. **j**, **k**, Representative images (**j**) and immunohistochemical staining (original magnification, ×40) (**k**) of tumours described in **i**. Data are mean ± s.d. (**c**–**e**, **g**, **h**) or mean ± s.e.m. (**i**). $^*P < 0.05$; $^{***}P < 0.001$.

(LDHA), which stimulates mitochondrial respiration in tumour cells[24], did not change the OCR nor promote senescence, whereas $BRAF^{V600E}$ did not affect LDHA protein levels (Supplementary Fig. 9a–f). Collectively, these results show that PDP2 and PDK1 are crucially required for the metabolic wiring associated with, and the execution of, OIS.

As OIS represents a rate-limiting step in oncogenic transformation[25], we next investigated whether PDK1 can act oncogenically. Melanocytes from the skin of neonatal *Tyr::CreER; Braf*^{CA} mice[14] were depleted of *p53* (also known as *TP53*), treated with tamoxifen to induce $BRAF^{V600E}$ expression and infected with a PDK1-encoding or control virus. Whereas transplanted control melanocytes expressing *Braf*^{V600E} and short hairpin RNA (shRNA) against *p53* (shp53) failed to form tumours, ectopic expression of PDK1 induced the formation of large tumours (Fig. 3i, j and Supplementary Fig. 10a). This was associated with robust PDH phosphorylation (Fig. 3k and Supplementary Fig. 10b).

These results raised the possibility that, conversely, PDK1 depletion acts cytostatically. Indeed, its silencing from non-transformed human cells induced proliferative arrest (Fig. 4a, b and Supplementary Fig. 11a, b). This was accompanied by suppression of the DNA replication-associated protein PCNA and induction of several senescence biomarkers and tumour suppressors (Fig. 4a, c and Supplementary Fig. 11b, c), thereby underscoring the importance of PDK1 for cellular senescence. PDK1-depleted cells also showed decreased PDH phosphorylation, enhanced PDH activity and an increased OCR (Fig. 4a, d, e). Thus, PDK1 depletion from non-transformed cells, including melanocytes, causes senescence. Unexpectedly, in a panel of human BRAF mutant melanoma cell lines, this cell cycle arrest was followed by cell death a few days later (Fig. 4f and Supplementary Fig. 11d, e). This was not caused by differences in the extent of PDK1 knockdown, PDH phosphorylation or activity, nor OCR (Supplementary Figs 11f–h and 12a, b).

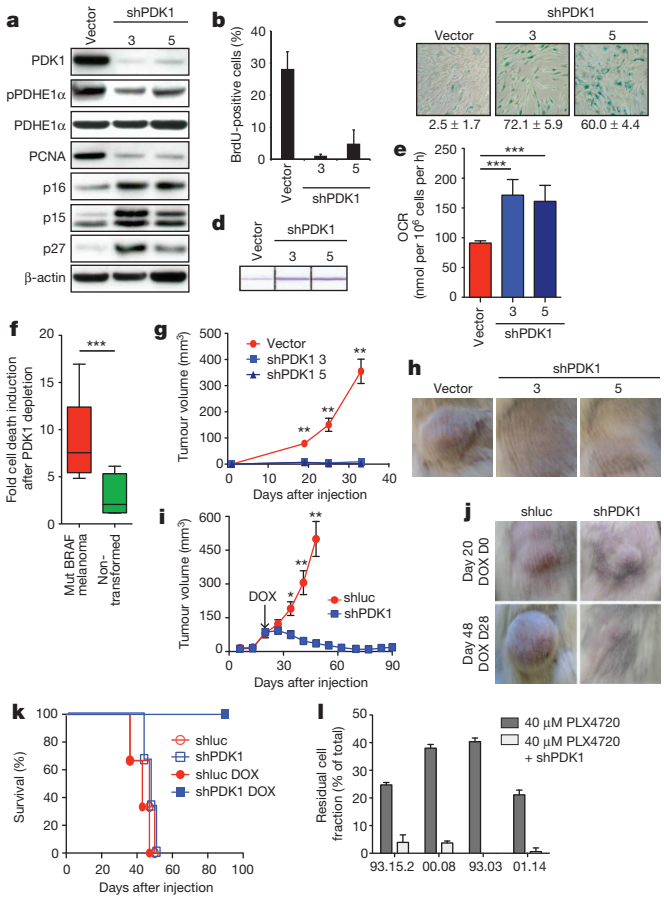


Figure 4 | PDK1 depletion causes melanoma regression and eradicates subpopulations resistant to targeted $BRAF^{V600E}$ inhibition. **a**–**e**, HDFs expressing empty vector or shPDK1 were analysed by immunoblotting (**a**) and for BrdU incorporation (**b**), SA-β-gal activity (**c**), PDH activity (**d**) and the OCR (**e**); $n = 3$. **f**, Cell death induction in melanoma ($n = 8$) and non-transformed cells ($n = 4$) after PDK1 depletion. **g**–**j**, Growth curves (**g**, **i**) and representative images (**h**, **j**) of tumours after constitutive (**g**, **h**) or DOX-inducible (**i**, **j**) PDK1 depletion; $n = 6$. **k**, Kaplan–Meier survival curve of mice from **i**. **l**, Residual cell fraction of control or PDK1-depleted melanoma cells treated with 40 μM PLX4720. Data are mean ± s.d. (**c**, **e**, **f**, **l**) or mean ± s.e.m. (**g**, **i**). $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

This led to the intriguing possibility that PDK1 depletion negatively affects melanoma growth. Indeed, PDK1-depleted melanoma cell lines (which were viable at the time of inoculation) almost completely failed to produce tumours in immunocompromised mice (Fig. 4g, h and Supplementary Fig. 13a–d). A few small lesions that did develop had invariably lost PDK1 knockdown (Supplementary Fig. 13b, e), indicating that PDK1 is essential for melanoma outgrowth *in vivo*. Because, clinically, it would be more relevant to assess the role of PDK1 in tumour maintenance and progression than in initiation, we generated a doxycycline (DOX)-inducible shRNA system. DOX administration suppressed PDH phosphorylation and concomitantly caused melanoma cell death *in vitro* (Supplementary Fig. 14a–c). In mice, uninduced shPDK1 cells produced tumours indistinguishably from control cells (Supplementary Fig. 14d). By contrast, when DOX was administered starting from the time of injection, PDK1-depleted cells failed to produce tumours (Supplementary Fig. 14e). Most importantly, when DOX was administered after melanomas had established, PDK1 depletion led to near-complete tumour regression, thereby greatly extending animal survival (Fig. 4i–k). These results demonstrate that PDK1 is not only required for tumour initiation, but also for tumour maintenance and progression, indicating that it may be beneficial to target this metabolic kinase for therapeutic intervention of BRAF mutant melanoma.

Finally, we examined whether PDK1 depletion sensitizes BRAF mutant melanoma cells to treatment with PLX4720 (a preclinical analogue of vemurafenib, a specific clinical $BRAF^{V600E}$ inhibitor[26,27]). Dose–response curves of four $BRAF^{V600E}$ melanoma cell lines that are sensitive to PLX4720 (>90% cell death after treatment with 40 μM PLX4720) and another four that are partially resistant (>20% cells surviving treatment with 40 μM PLX4720) revealed that PDK1 depletion strongly sensitized all cell lines to BRAF inhibition (Supplementary Fig. 15a, b). Remarkably, PDK1 depletion specifically eliminated melanoma subpopulations that resisted high PLX4720 concentrations (Fig. 4l and Supplementary Fig. 15b). Together, these observations indicate that PDK1 depletion synergizes with targeted BRAF inhibition to kill melanoma cells.

In conclusion, by metabolic profiling and subsequent functional perturbations of a key metabolic axis, we unveil that PDH, a gatekeeper linking glycolysis to oxidative metabolism, acts as a key regulator of OIS (Supplementary Fig. 1). The observation that PDH activity is induced during OIS and normalizes after OIS abrogation highlights this enzyme as a potential barrier against malignant transformation. In agreement, high PDK1 expression drives PDH phosphorylation and promotes melanoma growth. Conversely, PDK1 depletion is highly toxic to melanoma cells. Indeed, the regression of established mutant BRAF melanomas, plus the synergistic toxicity with targeted BRAF inhibition after PDK1 depletion raise the possibility that this metabolic kinase represents an attractive combinatorial target for therapeutic intervention of BRAF mutant metastatic melanoma.

METHODS SUMMARY

The HDF cell line TIG3 expressing the ectopic receptor and human telomerase reverse transcriptase (hTERT) (or its derivative expressing $shp16^{INK4A}$), the human retinal pigment epithelial cell line RPE1, the human prostate cell line PNT1A and all BRAF mutant melanoma cell lines (A0, mel:00.08, 01.14, 04.01, 04.07, 06.04, 07.16, 93.03, 93.15.2, 634, SK-MEL-23 and SK-MEL-28) were maintained in DMEM, supplemented with 9% FBS (PAA), 2 mM glutamine, 100 U ml^{-1} penicillin and 0.1 mg ml^{-1} streptomycin (GIBCO). Melanocytes were maintained as described

previously[12]. Lentiviral and retroviral infections were performed using HEK293T cells and Phoenix cells, respectively, as producers of viral supernatants. For senescence experiments, HDFs were infected with shRNA-encoding or protein-coding retro- or lentivirus, selected pharmacologically (puromycin or blasticidin) and subsequently infected with $BRAF^{V600E}$-encoding or control virus. After selection, cells were seeded for the cell proliferation assay, BrdU incorporation assay or SA-β-gal activity, and analysed. The OCR was measured using a XF24 extracellular flux analyser (Seahorse Bioscience). Exchange rate (uptake or secretion) measurements of key metabolites and stable isotope labelling were performed as described previously[28–30]. ROS production was measured with CellROX Deep Red Reagent from Invitrogen. The GSH/GSSG ratio was determined using the GSH/GSSG-Glo assay (Promega). PDH activity was measured using the DipStick assay kit (MitoSciences). Cell death induction was measured by the trypan blue exclusion assay. Transcripts and/or protein levels of the indicated genes were determined by quantitative reverse transcription PCR (qRT–PCR) and immunoblotting or immunohistochemistry, respectively. Details are described in the Methods.

Full Methods and any associated references are available in the online version of the paper.

Received 28 February 2012; accepted 4 April 2013.
Published online 19 May 2013.

1. Campisi, J. Suppressing cancer: the importance of being senescent. *Science* **309,** 886–887 (2005).
2. Collado, M. & Serrano, M. Senescence in tumours: evidence from mice and humans. *Nature Rev. Cancer* **10,** 51–57 (2010).
3. Vredeveld, L. C. W. *et al.* Abrogation of $BRAF^{V600E}$-induced senescence by PI3K pathway activation contributes to melanomagenesis. *Genes Dev.* **26,** 1055–1069 (2012).
4. Kuilman, T., Michaloglou, C., Mooi, W. J. & Peeper, D. S. The essence of senescence. *Genes Dev.* **24,** 2463–2479 (2010).
5. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432,** 307–315 (2004).
6. Adams, P. D. Healing and hurting: molecular mechanisms, functions, and pathologies of cellular senescence. *Mol. Cell* **36,** 2–14 (2009).
7. DeBerardinis, R. J., Sayed, N., Ditsworth, D. & Thompson, C. B. Brick by brick: metabolism and tumor cell growth. *Curr. Opin. Genet. Dev.* **18,** 54–61 (2008).
8. Tennant, D. A., Durán, R. V. & Gottlieb, E. Targeting metabolic transformation for cancer therapy. *Nature Rev. Cancer* **10,** 267–277 (2010).
9. Wellen, K. E. & Thompson, C. B. Cellular metabolic stress: considering how cells respond to nutrient excess. *Mol. Cell* **40,** 323–332 (2010).
10. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324,** 1029–1033 (2009).
11. Campisi, J. Replicative senescence: an old lives' tale? *Cell* **84,** 497–500 (1996).
12. Michaloglou, C. *et al.* $BRAF^{E600}$-associated senescence-like cell cycle arrest of human naevi. *Nature* **436,** 720–724 (2005).
13. Kuilman, T. *et al.* Oncogene-induced senescence relayed by an interleukin-dependent inflammatory network. *Cell* **133,** 1019–1031 (2008).
14. Dankort, D. *et al.* $Braf^{V600E}$ cooperates with *Pten* loss to induce metastatic melanoma. *Nature Genet.* **41,** 544–552 (2009).
15. Dhomen, N. *et al.* Oncogenic Braf induces melanocyte senescence and melanoma in mice. *Cancer Cell* **15,** 294–303 (2009).
16. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417,** 949–954 (2002).
17. Wieland, O. H. The mammalian pyruvate dehydrogenase complex: structure and regulation. *Rev. Physiol. Biochem. Pharmacol.* **96,** 123–170 (1983).
18. Patel, M. S. & Roche, T. E. Molecular biology and biochemistry of pyruvate dehydrogenase complexes. *FASEB J.* **4,** 3224–3233 (1990).
19. Kolobova, E., Tuganova, A., Boulatnikov, I. & Popov, K. M. Regulation of pyruvate dehydrogenase activity through phosphorylation at multiple sites. *Biochem. J.* **358,** 69–77 (2001).
20. Roche, T. E. *et al.* Distinct regulatory properties of pyruvate dehydrogenase kinase and phosphatase isoforms. *Prog. Nucleic Acid Res. Mol. Biol.* **70,** 33–75 (2001).
21. Holness, M. J. & Sugden, M. C. Regulation of pyruvate dehydrogenase complex activity by reversible phosphorylation. *Biochem. Soc. Trans.* **31,** 1143–1151 (2003).
22. Patel, M. S. & Korotchkina, L. G. Regulation of the pyruvate dehydrogenase complex. *Biochem. Soc. Trans.* **34,** 217–222 (2006).
23. Lemons, J. M. S. *et al.* Quiescent fibroblasts exhibit high metabolic activity. *PLoS Biol.* **8,** e1000514 (2010).
24. Fantin, V. R., St-Pierre, J. & Leder, P. Attenuation of LDH-A expression uncovers a link between glycolysis, mitochondrial physiology, and tumor maintenance. *Cancer Cell* **9,** 425–434 (2006).
25. Mooi, W. J. & Peeper, D. S. Oncogene-induced cell senescence–halting on the road to cancer. *N. Engl. J. Med.* **355,** 1037–1046 (2006).
26. Tsai, J. *et al.* Discovery of a selective inhibitor of oncogenic B-Raf kinase with potent antimelanoma activity. *Proc. Natl Acad. Sci. USA* **105,** 3041–3046 (2008).
27. Flaherty, K. T., Yasothan, U. & Kirkpatrick, P. Vemurafenib. *Nature Rev. Drug Discov.* **10,** 811–812 (2011).
28. Frezza, C. *et al.* Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature* **477,** 225–228 (2011).
29. Frezza, C. *et al.* Metabolic profiling of hypoxic cells revealed a catabolic signature required for cell survival. *PLoS ONE* **6,** e24411 (2011).
30. Chaneton, B. *et al.* Serine is a natural ligand and allosteric activator of pyruvate kinase M2. *Nature* **491,** 458–462 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J.-Y. Song for pathological analysis, M. McMahon for providing $Braf^{CA}$ mice, C. Vogel for sharing cell lines, R. van Amerongen for critical reading of the manuscript, and all members of the Gottlieb and Peeper laboratories for their input. This work was supported by Cancer Research UK, Spanish Government-EU-FEDER (grants SAF2011-25726 and ISCIII-RTICC-RD6/0020/0046) and ICREA-Academia to M.C., Israel Cancer Research Foundation and Israel Science Foundation to T.S., a Vici grant from the Netherlands Organization for Scientific Research (NWO) and a Queen Wilhelmina Award grant from the Dutch Cancer Society (KWF Kankerbestrijding) to D.S.P.

Author Contributions J.K., E.G. and D.S.P. conceived the project, analysed the data and wrote the manuscript. J.K. performed all *in vitro* experiments and carried out the *in vivo* experiments together with K.M. J.K., K.M. and B.C. performed metabolic experiments. L.Z. and G.M. performed LC–MS analyses. S.H.B. and E.M.E.V. provided low passage melanoma cell lines. V.A.S., M.C. and T.S. helped with metabolic analyses. All authors discussed the results and commented on the manuscript. E.G. and D.S.P. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.G. (e.gottlieb@beatson.gla.ac.uk) or D.S.P. (d.peeper@nki.nl).

METHODS

Cell culture, viral transduction and senescence induction. The HDF cell line TIG3 expressing the ectopic receptor and human telomerase reverse transcriptase (hTERT) (or its derivative expressing shp16^{INK4A}), the human retinal pigment epithelial cell line RPE1, the human prostate cell line PNT1A and all BRAF mutant melanoma cell lines (A0, mel:00.08, 01.14, 04.01, 04.07, 06.04, 07.16, 93.03, 93.15.2, 634, SK-MEL-23 and SK-MEL-28) were maintained in DMEM, supplemented with 9% FBS (PAA), 2 mM glutamine, 100 U ml^{-1} penicillin and 0.1 mg ml^{-1} streptomycin (GIBCO). Melanocytes were maintained as described previously[12]. Lentiviral and retroviral infections were performed using HEK293T cells and Phoenix cells, respectively, as producers of viral supernatants. For senescence experiments, HDFs were infected with shRNA-encoding or protein-coding retro- or lentivirus, selected pharmacologically (puromycin or blasticidin) and subsequently infected with BRAF^{V600E}-encoding or control virus. After selection, cells were seeded for the cell proliferation assay, BrdU incorporation assay or SA-β-gal activity, and analysed.

Measurement of metabolites by LC–MS. Stable isotope labelling was performed as described previously[28]. Two-million HDFs were plated onto 10-cm dishes and cultured in standard medium for 24 h. For stable isotope labelling analysis, the medium was replaced with 4.5 mM [U-^{13}C]glucose (Cambridge Isotope). After incubation for the indicated time, cells and media were collected. For extracellular metabolite analysis, 200 µl of growth medium from cell culture were added to 600 µl of acetonitrile for deproteinization. Samples were vortexed for 10 min and centrifuged for 10 min at 16,000*g* at 4 °C. The supernatant was stored for subsequent liquid chromatograph–mass spectrometry (LC–MS) analysis. For intracellular metabolite analysis, cells were lysed with a solution consisting of 50% methanol and 30% acetonitrile in water in dry ice methanol (−80 °C) and quickly scraped from the plate. The insoluble material was immediately pelleted in a cooled centrifuge (4 °C) for 10 min at 16,000*g*, and the supernatant was collected for subsequent LC–MS analysis. For [U-^{13}C]pyruvate labelling analysis, samples were processed as for the ^{13}C-labelled glucose labelling analysis, except that the medium was not replaced, but spiked with 0.11 mg ml^{-1} [$^{13}C_3$]sodium pyruvate (Cambridge Isotope). LC–MS analysis was carried out as described previously[29]. Mass spectrometry data were analysed by LCquan (Thermo Scientific), and quantifications of intracellular and extracellular metabolites were performed by the standard dilution method as described previously[30].

Cell proliferation assay. Cells were seeded into a six-well plate (at densities of 2×10^5, 4×10^5 or 6×10^5 cells) and selected pharmacologically. Fixation and staining with crystal violet was performed 9–13 days after the BRAF^{V600E}-encoding or control virus infection, or 6–9 days after shPDK1-encoding or control virus infection. Images of cell proliferation assays reflect representative results of at least three independent experiments.

BrdU incorporation assay. BrdU labelling was carried out for 3 h followed by fixation. Incorporated BrdU was detected by immunostaining as described previously[31] and by FACS analysis. Results are represented as mean and s.d. of at least three independent experiments.

Analysis of SA-β-gal activity. SA-β-gal was stained using the Senescence Associated β-Galactosidase Staining kit (Cell Signaling) at pH 6, according to the manufacturer's protocol. Images reflect representative results of at least three independent experiments.

Trypan blue exclusion assay. Cells were brought into suspension using trypsin, centrifuged and resuspended in a small volume of culture medium. Trypan blue (Sigma) was added to the cell suspension (dilution factor = 2) and stained cells were counted as dead. The number of dead cells was quantified, and the values were expressed as the fold change over control. Results are represented as mean and s.d. of at least three independent experiments.

Measurement of redox stress. ROS production was measured with CellROX Deep Red Reagent from Invitrogen. Cells were incubated at 37 °C for 30 min in DMEM supplemented with 9% FBS (PAA) and containing 2.5 µM CellROX Deep Red Reagent. Cells were then washed twice in PBS, treated with trypsin and resuspended in PBS supplemented with 50% FCS. Fluorescence was immediately measured using FACS analysis.

The GSH/GSSG ratio was determined using the GSH/GSSG-Glo assay (Promega) according to the manufacturer's protocol. Results are represented as mean and s.d. of at least three independent experiments.

Plasmids. pMSCV-blast-BRAF^{V600E} and pMSCV-blast were previously described[13]. For the overexpression of PDK1, *PDK1* derived from complementary DNA of a normal human skin sample was PCR-amplified and cloned into pLZRS-IRES-puro or FG12-eGFP. Empty pLZRS-IRES-puro or FG12-eGFP was used as a control.

shRNAs. Retroviral knockdown constructs were described previously[12,13]. Lentiviral constitutive knockdown constructs were purchased from Sigma-Aldrich in pLKO.1 backbone or cloned (shp53) into KH1eGFP backbone (a gift from M. Soengas): shPDP2-1 (clone TRCN0000036739), 5′-CCTTGAAAGCAGAGTCCCAAA-3′; shPDP2-4 (TRCN0000036742) 5′-GCTGAAGTGGAGTAAAGAGTT-3′; shPDK1-3 (TRCN0000006261) 5′-GCTCTGTCAACAGACTCAATA-3′; shPDK1-5 (TRCN0000006263) 5′-CCAGGGTGTGATTGAATACAA-3′; shp53 mouse, 5′- GTACATGTGTAATAG CTCC-3′.

As negative controls pRS-puro, pLKO.1-puro or KH1eGFP without insert were used. Lentiviral-inducible knockdown plasmid pLKO.1-Tet-On was purchased from Addgene. shRNA targeting PDK1 were re-cloned from the constitutive shPDK1-3 knockdown construct. As negative control pLKO.1-Tet-On with shRNA targeting luciferase was used: shluc, 5′-CGCTGAGTACTTCGAAATGTC-3′.

qRT–PCR. Total RNA was DNase-treated with RQ1 RNase-Free DNase (Promega). Reverse transcription was performed with SuperScript II First Strand Kit (Invitrogen). qRT–PCR was performed with the SYBR Green PCR Master Mix (Applied Biosystems) on an ABI PRISM 7700 Sequence detection system.

IL6, *IL8*, *CEBPB* and *RPL13* (standard) primer sequence were described previously[13]. Other primer sets used were as follows: *PDK1*, 5′-CCAAGACCTCGTGTTGAGA CC-3′ and 5′-AATACAGCTTCAGGTCTCCTTGG-3′; *PDK2*, 5′-GAGCCTCC TGGACATCATGG-3′ and 5′-TACTCAAGCACGCCTTGTGC-3′; *PDK3*, 5′-A CTGTATTCCATGGAAGGAGTGG-3′ and 5′-CTCCAATCATCGGCTTCAG G-3′; *PDK4*, 5′-AACTGTGATGTGGTAGCAGTGG-3′ and 5′-GATGTGAATT GGTTGGTCTGG-3′; *PDP2*, 5′-ACCACCTCCGTGTCTATTGG-3′ and 5′-CC AGCGAGATGTCAGAATCC-3′; *NQO1*, 5′-CAGCTCACCGAGAGCCTAGT-3′ and 5′-GAGTGAGCCAGTACGATCAGTG-3′; *GCLC*, 5′-ATGCCATGGGA TTTGGAAT-3′ and 5′-AGATATACTGCAGGCTTGGAATG-3′; *GSTA4*, 5′-AG TTGTACAAGTTGCAGGATGG-3′ and 5′-CAATTTCAACCATGGGCACT-3′; *GSTM4*, 5′-TCATCTCCCGCTTTGAGG-3′ and 5′-CAGACAGCCACCCTTGT GTA-3′; *HMOX1*, 5′-GGGTGATAGAAGAGGCCAAGA-3′ and 5′-AGCTCCTG CAACTCCTCAAA-3′; *SOD1*, 5′-TCATCAATTTCGAGCAGAAGG-3′ and 5′-CA GGCCTTCAGTCAGTCCTTT-3′; *SOD2*, 5′-CTGGACAAACCTCAGCCCTA-3′ and 5′-TGATGGCTTCCAGCAACTC-3′.

Except for *CEBPB*, all primer pairs span exon–exon borders. *RPL13* was used as a control. For analysis, the ΔC_T method was applied. Data are represented as mean ± s.d. of three or more independent experiments.

Antibodies. Antibodies used for immunoblotting/immunohistochemistry were β-actin (AC-74; A5316; Sigma), BRAF (sc-5284; Santa Cruz), Hsp90 (4874; Cell Signaling), LDHA (2012; Cell Signaling), MEK1/2 (L38C12; 4694; Cell Signaling), phospho-MEK1/2 (Ser 217/221) ((41G9); 9154; Cell Signaling), PCNA ((PC10); sc-56; Santa Cruz), PDHE1α ((9H9AF5); 459400; Invitrogen), phospho-PDHE1α (Ser 293) (AP1062; Calbiochem), PDK1 (KAP-PK112; Stressgene), PDP2 (HPA01995; Sigma), p16^{INK4A} ((JC3); sc-56330; Santa Cruz), p27^{Kip1} (610241; BD Transduction Laboratories), and p15^{INK4B} (sc-612; Santa Cruz).

Immunohistochemistry. Formalin-fixed paraffin-embedded tissue samples were stained according to common procedures. For antigen retrieval, samples were incubated in 20 µg ml^{-1} proteinase K (Z0622, Sigma for PDK1), in citrate buffer (Biogenex, for PDHE1α) or in Tris-EDTA, pH 9.0 (for phosphor-PDHE1α (Ser 293)). Sections were counterstained with haematoxylin.

Drug treatment. For dose–response curves, melanoma cells were treated with the indicated concentrations of PLX4720 (Selleck) for 3 days. Cell viability was determined with the Cell Titer Blue assay (Promega) and fluorescence was measured with a TECAN infinite scanner. Results represent at least three independent experiments.

***In vivo* assays.** All mouse experiments were performed according to a protocol approved by the Institutional Animal Experiment Ethics Committee. Experiments were repeated at least twice.

Murine melanocyte derivation, culture and tumour growth *in vivo*. Melanoytes were derived from neonatal skin of *Tyr::CreER;Braf^{CA}* mice[14], as described previously[32]. Primary melanocyte cultures were prepared on a mitomycin-treated XB2 (immortal murine keratinocytes) feeder cell layer for one passage only. Cells were grown in RPMI medium supplemented with 5% FCS, 200 nM 12-*O*-tetradecanoyl phorbol 13-acetate, 200 pM cholera toxin, 10 ng ml^{-1} recombinant stem cell factor (SCF; R&D Systems) and 100 nM endothelin 3 (Bachem) at 37 °C and under low oxygen conditions (5% CO_2 and 3% O_2). Primary melanocytes were transduced with retro- or lentivirus in the presence of 2 µg ml^{-1} polybrene overnight. The transduction with lentivirus delivering shRNA targeting p53 and the PDK1 expression vector were done on consecutive nights. The transduction of virus allowing PDK1 expression was repeated two to four times. In addition, melanocytes were treated with 0.2 µM 4-hydroxytamoxifen for at least 9 days to induce the expression of BRAF^{V600E} by switching from the conditional to the mutated allele. One week after the last transduction, 1×10^6–1.5×10^6 melanocytes (dependent on the experiment) were subcutaneously injected with 50% basement membrane Matrigel (growth factors reduced) into both flanks of NSG mice (NOD *scid* IL2 receptor gamma chain knockout mice). Tumour volume was determined by measurement of two dimensions and calculation with the following formula: $V = 4/3 * \pi * a^2 * b$, in which a is the shorter and b is the longer dimension. BRAF allele rearrangements

in tumours were detected by PCR as previously described[33]. Tumours were analysed by immunohistochemistry and immunoblot.

Xenograft experiments. NOD/*scid* mice were subcutaneously injected with 0.5×10^6 cells into both flanks. For DOX-inducible shRNA xenograft experiments, mice were exposed to 2 mg ml^{-1} DOX administered in 5% sucrose-containing drinking water, either directly after injection or when tumours reached a volume of 100 mm^3. Mice were inspected twice a week and euthanized by CO_2 when tumours reached the volume of 500 mm^3 (1,000 mm^3 per mouse). Tumour volume was determined as above.

Measurement of OCR. Basal OCR was measured using the XF24 extracellular flux analyser (Seahorse Bioscience). At the end of the experiment, 1 mmol l^{-1} antimycin A was added to measure mitochondria-independent oxygen consumption. Each cycle of measurement consisted of 3 min mixing, 3 min waiting and 4 min measuring. OCR was normalized to the cell number calculated at the end of the experiments. To obtain the mitochondrial-dependent OCR, only the antimycin-sensitive respiration was used. Homogeneous plating of the cells and cell count were assessed by fixing the cells with trichloroacetic acid 10% for 1 h at 4 °C and then staining the fixed cells with a 0.47% solution of sulphorhodamine B (Sigma).

Measurement of PDH activity. PDH activity in cell lysates was measured using the DipStick assay kit from MitoSciences (MSP90). Cells were lysed in the sample buffer provided by the manufacturer, followed by centrifugation and measurement of the protein concentration with the BioRad Protein Assay. One-hundred-and-forty milligrams of protein lysate was loaded and PDH activity was measured according to the manufacturer's protocol.

Statistical analysis. Statistical analyses of metabolite exchange rates were done with *t*-tests. Analysis of all other data was done with a non-parametric two-tailed Mann–Whitney *U* test with a 95% confidence interval (Prism; GraphPad Software). $P < 0.05$ was considered significant.

31. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic *ras* provokes premature cell senescence associated with accumulation of p53 and $p16^{INK4a}$. *Cell* **88,** 593–602 (1997).
32. Sviderskaya, E. V. *et al.* Complementation of hypopigmentation in *p*-mutant (*pink-eyed dilution*) mouse melanocytes by normal human P cDNA, and defective complementation by OCA2 mutant sequences. *J. Invest. Dermatol.* **108,** 30–34 (1997).
33. Dankort, D. *et al.* A new mouse model to explore the initiation, progression, and therapy of $BRAF^{V600E}$-induced lung tumors. *Genes Dev.* **21,** 379–384 (2007).

LETTER

doi:10.1038/nature12240

Innate lymphoid cells regulate CD4$^+$ T-cell responses to intestinal commensal bacteria

Matthew R. Hepworth[1,2], Laurel A. Monticelli[2,3], Thomas C. Fung[1,2,3], Carly G. K. Ziegler[4], Stephanie Grunberg[3], Rohini Sinha[3], Adriana R. Mantegazza[5], Hak-Ling Ma[6], Alison Crawford[2,3], Jill M. Angelosanto[2,3], E. John Wherry[2,3], Pandelakis A. Koni[7], Frederic D. Bushman[3], Charles O. Elson[8], Gérard Eberl[9,10], David Artis[2,3,11] & Gregory F. Sonnenberg[1,2]

Innate lymphoid cells (ILCs) are a recently characterized family of immune cells that have critical roles in cytokine-mediated regulation of intestinal epithelial cell barrier integrity[1–10]. Alterations in ILC responses are associated with multiple chronic human diseases, including inflammatory bowel disease, implicating a role for ILCs in disease pathogenesis[3,8,11–13]. Owing to an inability to target ILCs selectively, experimental studies assessing ILC function have predominantly used mice lacking adaptive immune cells[1–10]. However, in lymphocyte-sufficient hosts ILCs are vastly outnumbered by CD4$^+$ T cells, which express similar profiles of effector cytokines. Therefore, the function of ILCs in the presence of adaptive immunity and their potential to influence adaptive immune cell responses remain unknown. To test this, we used genetic or antibody-mediated depletion strategies to target murine ILCs in the presence of an adaptive immune system. We show that loss of retinoic-acid-receptor-related orphan receptor-γt-positive (RORγt$^+$) ILCs was associated with dysregulated adaptive immune cell responses against commensal bacteria and low-grade systemic inflammation. Remarkably, ILC-mediated regulation of adaptive immune cells occurred independently of interleukin (IL)-17A, IL-22 or IL-23. Genome-wide transcriptional profiling and functional analyses revealed that RORγt$^+$ ILCs express major histocompatibility complex class II (MHCII) and can process and present antigen. However, rather than inducing T-cell proliferation, ILCs acted to limit commensal bacteria-specific CD4$^+$ T-cell responses. Consistent with this, selective deletion of MHCII in murine RORγt$^+$ ILCs resulted in dysregulated commensal bacteria-dependent CD4$^+$ T-cell responses that promoted spontaneous intestinal inflammation. These data identify that ILCs maintain intestinal homeostasis through MHCII-dependent interactions with CD4$^+$ T cells that limit pathological adaptive immune cell responses to commensal bacteria.

ILCs are a heterogeneous population of innate immune cells that can be grouped based on their expression of, and developmental requirements for, specific transcription factors and cytokines[1,8–10]. Group 1 ILCs depend on T-bet and express interferon (IFN)-γ, whereas group 2 ILCs depend on ROR-α and GATA3 and express IL-5, IL-13 and amphiregulin[1,8–10]. Group 3 ILCs critically depend on RORγt for their development and in response to IL-23 stimulation produce the effector cytokines IL-17A and IL-22, which directly regulate innate immunity, inflammation and anatomical containment of pathogenic and commensal bacteria in the intestine[1,8–10,14]. However, the function of group 3 ILCs in the presence of adaptive immunity, and whether ILCs can influence adaptive immune cell responses, is unknown. To test this, adaptive immune cell responses were examined in mice lacking RORγt (*Rorc*$^{gfp/gfp}$). In comparison to control mice, *Rorc*$^{gfp/gfp}$ mice exhibited significantly increased frequencies of peripheral proliferating Ki-67$^+$CD4$^+$ T cells and effector/effector memory CD44^{high}CD62L^{low}CD4$^+$ T cells (Fig. 1a) and developed splenomegaly (Fig. 1b, c), indicative of disrupted immune cell homeostasis. *Rorc*$^{gfp/gfp}$ mice also exhibited elevated levels of commensal bacteria-specific serum IgG (Fig. 1d), suggesting that commensal bacteria were promoting activation of adaptive immune cells in the absence of RORγt. Consistent with this, oral administration of antibiotics to *Rorc*$^{gfp/gfp}$ mice was associated with significantly reduced peripheral Ki-67$^+$CD4$^+$ T cells and CD44^{high}CD62L^{low}CD4$^+$ T cells, spleen size and weight and commensal bacteria-specific serum IgG (Fig. 1a–d). As T cells also express RORγt and RORγt-deficient mice exhibit several developmental abnormalities[6,15,16], we used CD90-disparate chimaeras to allow transient depletion of CD90.2$^+$ ILCs, but not CD90.1$^+$ T cells[3,7]. Depletion of ILCs in CD90-chimaeric mice with an anti-CD90.2 monoclonal antibody resulted in significantly increased frequencies of dysregulated CD4$^+$ T cells, increased spleen weight and elevated commensal bacteria-specific serum IgG responses (Supplementary Fig. 1a–d), suggesting a critical role for ILCs in the regulation of inflammatory adaptive immune cell responses to commensal bacteria. Unexpectedly, IL-22-, IL-17A- and IL-23-deficient mice did not exhibit altered CD4$^+$ T-cell responses, splenomegaly or elevated levels of commensal bacteria-specific serum IgG (Supplementary Fig. 2a–c). Furthermore, transient blockade of IL-22, IL-17A, IL-23 or IL-17RA in C57BL/6 mice also failed to exacerbate adaptive immune cell responses to commensal bacteria (Supplementary Fig. 2d–i), indicating that ILCs regulate adaptive immune cell responses independently of effector cytokines.

To identify the mechanisms by which RORγt$^+$ group 3 ILCs regulate commensal bacteria-responsive adaptive immune cells, genome-wide transcriptional profiles of RORγt$^+$ ILCs were compared to those of naive CD4$^+$ T cells (Fig. 1e, f). Analysis of the top differentially expressed transcripts in RORγt$^+$ group 3 ILCs revealed a significant enrichment for genes involved in pathways of 'haematopoietic or lymphoid organ development' and 'immune response' (Fig. 1e), consistent with previous analyses of RORγt$^+$ ILCs[17,18]. Notably, an additional pathway that was highly enriched in the transcriptional profile of group 3 ILCs was 'antigen processing and presentation of peptide antigen via MHCII' (Fig. 1e). Indeed, relative to naive CD4$^+$ T cells (Fig. 1f) and previously published arrays of *in-vitro*-generated T-helper 17 cells[19] (Supplementary Fig. 3), group 3 ILCs were highly enriched in transcripts involved in MHCII antigen processing and presentation pathways, such as *Cd74*, *H2-DMb2*, *H2-DMa*, *H2-Ab1* and *H2-Aa*. Consistent with these transcriptional analyses, MHCII

[1]Division of Gastroenterology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [2]Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [3]Department of Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [4]ImmunoDynamics Group, Programs in Computational Biology and Immunology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [5]Department of Pathology and Laboratory Medicine, and Department of Physiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [6]Inflammation and Immunology Research Unit, Biotherapeutics Research and Development, Pfizer Worldwide R&D, Cambridge, Massachusetts 02140, USA. [7]Cancer Immunology, Inflammation & Tolerance Program, Georgia Health Sciences University Cancer Center, Augusta, Georgia 30912, USA. [8]Department of Medicine, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA. [9]Lymphoid Tissue Development Unit, Institute Pasteur, 75724 Paris, France. [10]Centre National de la Recherche Scientifique, URA 1961, 75724 Paris, France. [11]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

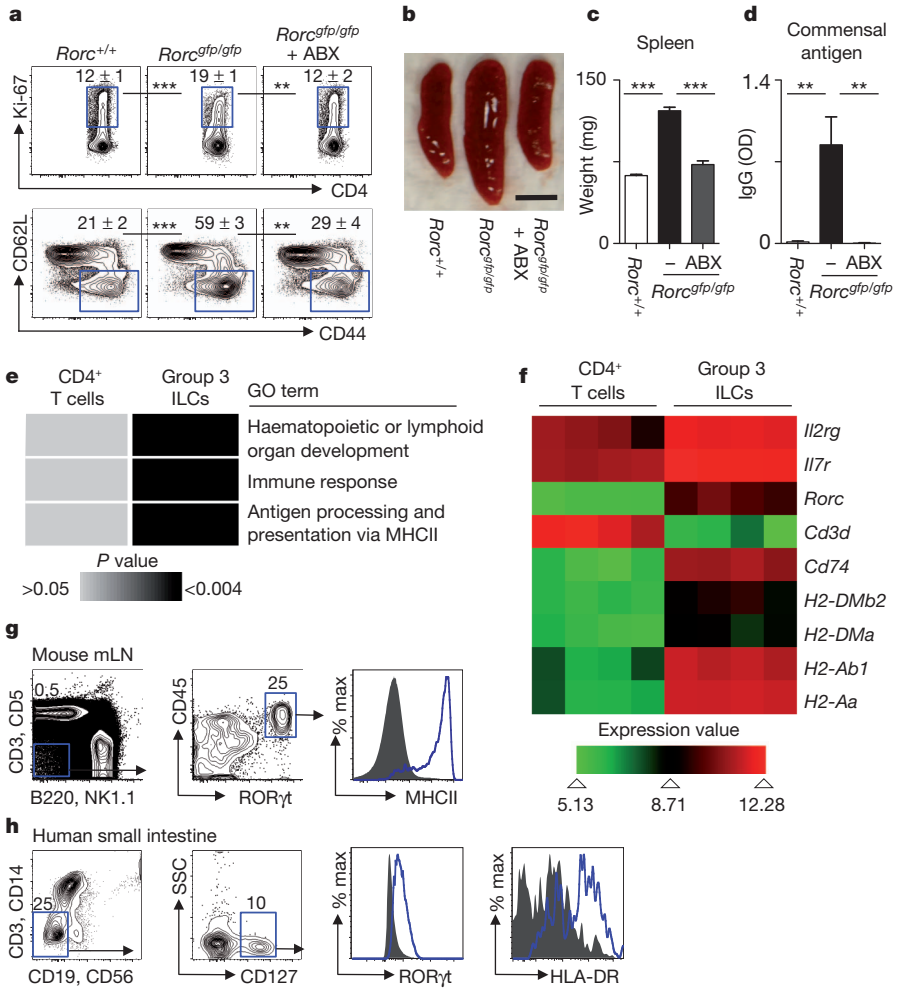


Figure 1 | RORγt$^+$ ILCs regulate adaptive immune cell responses to commensal bacteria and are enriched in MHCII-associated genes. **a–d**, Defined age- and sex-matched mouse strains were examined for the frequency of splenic Ki-67$^+$CD4$^+$ T cells (top) and CD44^{high}CD62L^{low}CD4$^+$ T cells (bottom) (**a**), spleen size (**b**), spleen weight (**c**) and relative absorbance values (OD) of serum IgG specific to commensal bacteria (**d**). Antibiotics (ABX) were administered in the drinking water from weaning until 6–8 weeks of age. Scale bar, 0.5 cm (**b**). Flow cytometry plots are gated on live CD4$^+$CD3$^+$ T cells (**a**). **e**, **f**, DAVID pathway analysis of GO terms enriched in the transcriptional profiles of naive CD4$^+$ T cells and group 3 RORγt$^+$ ILCs (**e**) and heat map of selected lymphoid-associated and MHCII-associated gene transcripts (**f**). **g**, **h**, Gating strategy for ILCs and expression of RORγt and MHCII in ILCs from the mesenteric lymph node (mLN) of naive RORγt–eGFP reporter mice (**g**) and the small intestine of healthy humans (**h**). Blue line, ILCs; grey fill, negative control population. Data are representative of three independent experiments containing 3–5 mice per group or four human donors. Results are shown as the means ± s.e.m. $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed Student's *t*-test).

protein was detected on gated lineage$^-$CD45$^+$RORγt$^+$ ILCs from the mesenteric lymph node of naive RORγt–eGFP reporter mice (Fig. 1g). Critically, MHCII protein was also identified on gated lineage$^-$CD127$^+$RORγt$^+$ ILCs from the small intestine of healthy humans (Fig. 1h). Collectively, these results indicate that group 3 RORγt$^+$ ILCs in the intestinal and lymphoid tissues of healthy mice and humans express MHCII.

To interrogate whether MHCII expression was restricted to group 3 ILCs, total lineage$^-$CD45$^+$CD90.2$^+$ ILCs in the murine small intestine were subdivided into group 1, 2 and 3 ILCs by expression of their defining transcription factors (Fig. 2a, b and Supplementary Fig. 4a–e). Group 1 ILCs (RORγt$^-$T-bet$^+$NKp46$^{+/-}$) were found to lack MHCII expression (Supplementary Fig. 4d–f), whereas group 2 ILCs (GATA-3$^+$) expressed intermediate levels of MHCII (Supplementary Fig. 4c). Microarray analyses confirmed the enrichment of MHCII-associated genes in group 3 ILCs versus previously published arrays of group 1 ILCs, such as natural killer (NK) cells[20] (Supplementary Fig. 5a) and group 2 ILCs[17] (Supplementary Fig. 5b). Significant heterogeneity exists within RORγt$^+$ group 3 ILCs and three subgroups can be identified on the basis of expression of NKp46 and T-bet (Fig. 2a)[9,10,21]. MHCII was found to be highly expressed on RORγt$^+$ ILCs that lacked expression of both T-bet and NKp46, whereas minimal expression of MHCII was observed on RORγt$^+$T-bet$^+$NKp46$^-$ and RORγt$^+$T-bet$^+$NKp46$^+$ ILC subsets (Fig. 2b) isolated from the small intestine of naive mice. Consistent with this, previously published microarray data profiling ILCs based on NKp46 and RORγt expression[18] also revealed an enrichment of MHCII-associated genes in RORγt$^+$ ILCs that lacked NKp46 and T-bet expression (Supplementary Fig. 5c). Furthermore, MHCII$^+$RORγt$^+$ ILCs were found in lymphoid tissues at steady state (Supplementary Fig. 6), exhibited homogeneous expression of CD127, CD90.2, CD25, CCR6, c-kit and CD44, and heterogeneous expression of Sca-1 and CD4 (Supplementary Fig. 7a) and produced IL-22, but not IL-17A or IFN-γ, in response to IL-23 stimulation (Supplementary Fig. 7b).

To interrogate the functional capacity of MHCII$^+$ ILCs, cells were sort-purified and cultured with DQ-ovalbumin (DQ-OVA), a self-quenching conjugate of ovalbumin that fluoresces upon proteolytic degradation. MHCII$^+$ ILCs exhibited an increase in fluorescence intensity comparable to CD11c$^+$MHCII$^+$ dendritic cells after incubation with DQ-OVA (Fig. 2c), indicative of an ability to acquire and degrade antigens. Sort-purified ILCs were also cultured with green fluorescent protein (GFP)-labelled E-alpha (Eα) protein and stained with an antibody specific for Eα-derived $E\alpha_{52-68}$ peptide bound to I–A^b molecules (Y-Ae). MHCII$^+$ ILCs incubated with GFP–Eα exhibited positive GFP fluorescence and staining for Y-Ae at levels comparable to those of CD11c$^+$MHCII$^+$ dendritic cells (Fig. 2d), demonstrating that ILCs can process exogenous protein and present peptide antigen in the context of MHCII. However, in contrast to OVA-pulsed dendritic cells that induced multiple rounds of ovalbumin-specific CD4$^+$ T-cell proliferation, OVA-pulsed ILCs failed to induce ovalbumin-specific CD4$^+$ T-cell proliferation (Fig. 2e). Consistent with this, MHCII$^+$RORγt$^+$ ILCs lacked expression of the classical co-stimulatory molecules CD40, CD80 and CD86, relative to dendritic cells (Fig. 2f). Antigen presentation in the absence of co-stimulatory molecules has been proposed to limit T-cell responses[22], suggesting that MHCII$^+$RORγt$^+$ ILCs may negatively regulate CD4$^+$

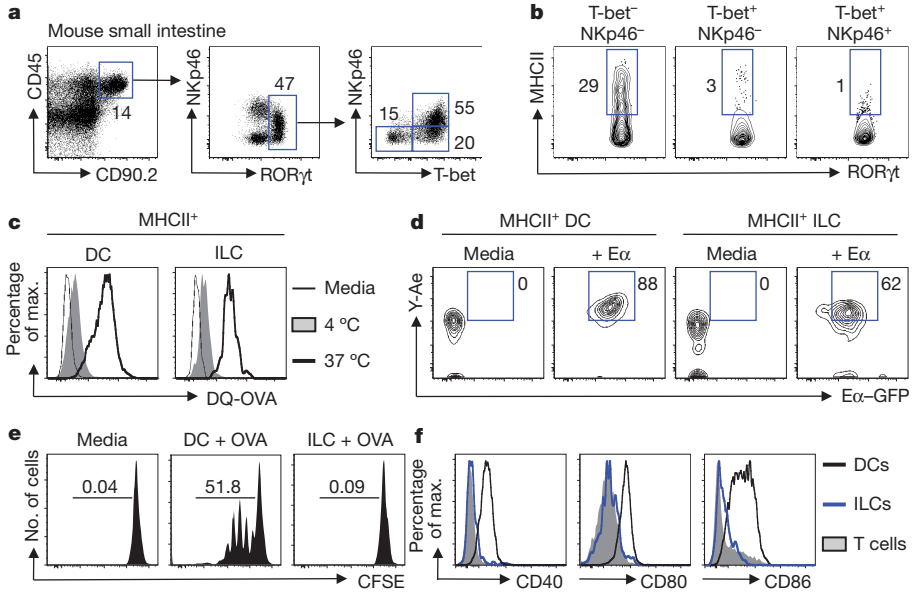


Figure 2 | T-bet$^-$NKp46$^-$RORγt$^+$ ILCs express MHCII and process and present antigen, but do not induce T-cell proliferation. **a**, **b**, Gating strategy (**a**) and expression of MHCII (**b**) in group 3 ILC subsets in the small intestine of naive mice. **c**, **d**, Sorted cell populations were cultured in the absence (thin line) or presence of DQ-OVA at 4 °C (shaded) or 37 °C (thick line) (**c**) or cultured in the absence (media) or presence of Eα–GFP protein and stained with Y-Ae antibody (**d**). **e**, Sort-purified CFSE-labelled CD4$^+$ T cells from OT-II mice were cultured in the presence of media alone or with OVA-pulsed dendritic cells (DCs) or OVA-pulsed ILCs. **f**, Expression of co-stimulatory molecules on dendritic cells (black line), ILCs (blue line) or T cells (shaded) from the mesenteric lymph nodes of naive mice. Data are representative of 2–3 independent experiments containing 2–5 mice per group or 2–3 *in vitro* replicates.

T-cell responses *in vivo*. To test this, ILCs were pulsed with the commensal bacteria-derived antigen CBir1 (ref. 23) and co-transferred with CBir1-specific transgenic T cells into naive congenic mice, before systemic challenge with peptide (Supplementary Fig. 8a). Mice that received a co-transfer of CBir1-specific T cells with peptide-pulsed ILCs exhibited a reduced population expansion of transferred T cells and decreased antigen-specific IFN-γ production relative to transfer of T cells alone (Supplementary Fig. 8b–d), suggesting that antigen presentation by ILCs limits CD4$^+$ T-cell responses *in vivo*.

To investigate further the ability of MHCII$^+$RORγt$^+$ ILCs to regulate adaptive immune cell responses, mice were generated with a RORγt$^+$ ILC-intrinsic deletion of MHCII (MHCII$^{\Delta ILC}$) by crossing

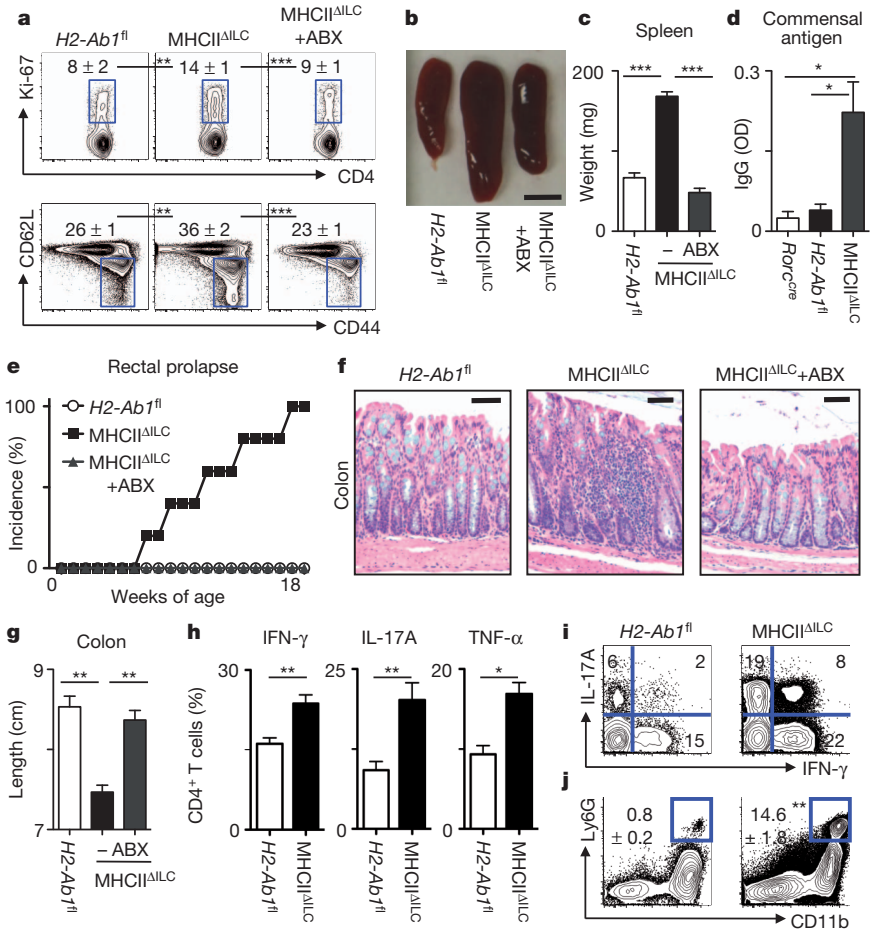


Figure 3 | Loss of RORγt$^+$ ILC-intrinsic MHCII expression results in commensal bacteria-dependent intestinal inflammation. **a**–**d**, Age- and sex-matched mice were examined for the frequency of splenic Ki-67$^+$CD4$^+$ T cells (top) and CD44^{high}CD62L^{low}CD4$^+$ T cells (bottom) (**a**), spleen size (**b**), spleen weight (**c**) and relative serum IgG specific to commensal bacteria (**d**). Scale bar, 0.5 cm (**b**). Flow cytometry plots are gated on live CD4$^+$CD3$^+$ T cells (**a**). **e**–**g**, Mice were examined for incidence of rectal prolapse (**e**), histological changes in haematoxylin and eosin stained sections of the terminal colon (**f**) and colon length (**g**). Scale bars, 25 μm (**f**). Antibiotics (ABX) were continuously administered in the drinking water of selected mice at weaning until 8–18 weeks of age. **h**–**j**, Frequency of total IFN-γ$^+$, IL-17A$^+$ and TNF-α$^+$ CD4$^+$ T cells (**h**) and IL-17A$^+$IFN-γ$^+$CD4$^+$ T cells (**i**) in the colons after a brief *ex vivo* stimulation and frequency of CD11b$^+$ Ly6G$^+$ neutrophils in colonic lamina propria (**j**). Data are representative of three independent experiments containing 3–5 mice per group. Results are shown as the means ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (two-tailed Student' *t*-test).

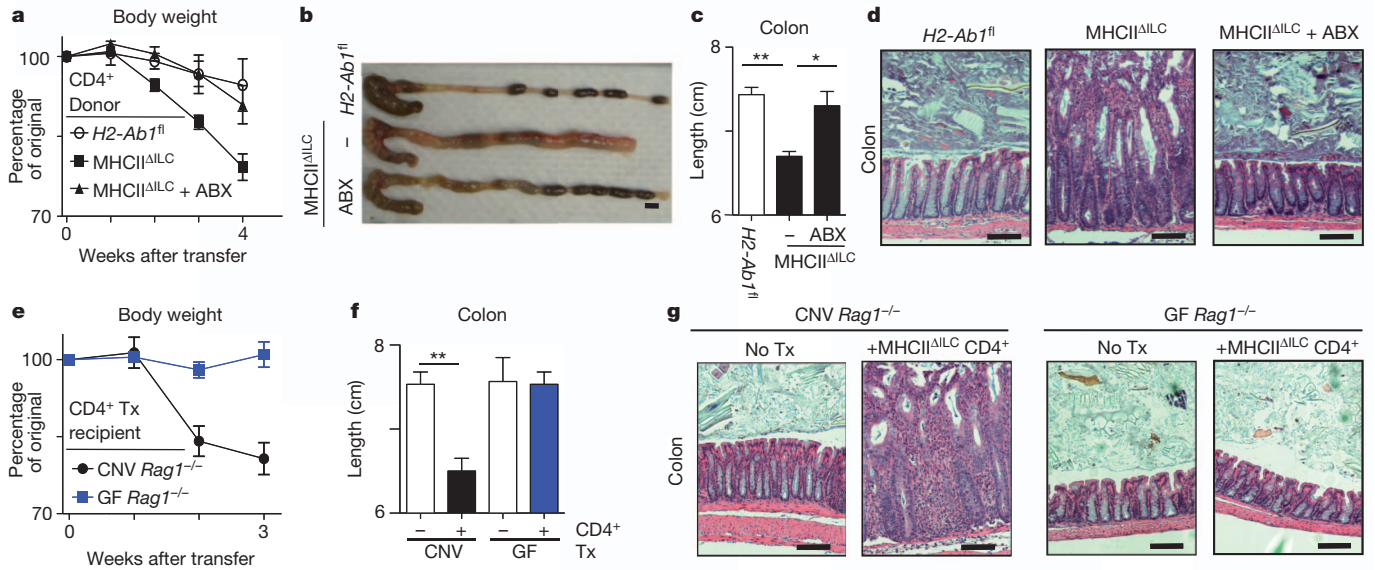


Figure 4 | RORγt$^+$ ILC-intrinsic MHCII regulates pathological CD4$^+$ T-cell responses to commensal bacteria. **a**–**g**, *Rag1*$^{-/-}$ mice received CD4$^+$ T cells sort-purified from defined donor mouse strains (**a**–**d**), or conventional (CNV) and germ-free (GF) *Rag1*$^{-/-}$ mice received sort-purified CD4$^+$ T cells from MHCII$^{\Delta ILC}$ mice via adoptive transfer (± Tx) (**e**–**g**). Recipients were examined for changes in weight (**a**, **e**), macroscopic colon pathology (**b**), colon length (**c**, **f**) and histological changes in the terminal colon (**d**, **g**). Scale bars, 0.5 cm (**b**) or 25 μm (**d**, **g**). In some experiments, antibiotics (ABX) were administered in the drinking water of donor mice from weaning (**a**–**d**). Data are representative of two independent experiments containing 3–5 mice per group. Results are shown as the means ± s.e.m. $*P < 0.05$, $**P < 0.01$ (two-tailed Student's *t*-test).

mice with a floxed *H2-Ab1* gene (*H2-Ab1*^{fl}) with mice expressing Cre recombinase under the control of the *Rorc* promoter (*Rorc*^{cre}) (Supplementary Fig. 9a). Given that *Rorc* is expressed only by T cells and ILCs[15,16] and that murine T cells do not express MHCII[24], this permitted selective genetic deletion of MHCII in RORγt$^+$ ILCs in the presence of an intact adaptive immune system. Consistent with this, MHCII$^{\Delta ILC}$ mice exhibited a selective loss of MHCII expression on RORγt$^+$ ILCs, whereas B cells, dendritic cells and macrophages retained comparable expression levels of MHCII relative to control *H2-Ab1*^{fl} mice (Supplementary Fig. 9b). MHCII$^{\Delta ILC}$ mice also had comparable peripheral numbers of lymph nodes and Peyer's patches, frequencies of ILCs and production of ILC-derived IL-22 (Supplementary Fig. 9c, d) as compared to control mice. However, MHCII$^{\Delta ILC}$ mice did exhibit significantly increased frequencies of peripheral Ki-67$^+$CD4$^+$ T cells and CD44^{high}CD62L^{low}CD4$^+$ T cells, increased spleen size and weight and a significant increase in commensal bacteria-specific serum IgG, which were abrogated upon oral administration of antibiotics (Fig. 3a–d). Thus, loss of ILC-intrinsic MHCII expression recapitulated the phenotype observed after genetic or antibody-mediated depletion of ILCs (Fig. 1a–d and Supplementary Fig. 1a–d) identifying a critical role for MHCII$^+$ ILCs in regulating T-cell responses to commensal bacteria.

Inappropriate host inflammatory responses to commensal bacteria are associated with the pathogenesis and progression of numerous chronic human diseases[3,11–13], therefore MHCII$^{\Delta ILC}$ mice were examined at various ages for signs of inflammation. MHCII$^{\Delta ILC}$ mice were observed to develop rectal prolapse, beginning at approximately 8 weeks of age and reaching a 100% incidence by 18 weeks of age (Fig. 3e). Further examination revealed that MHCII$^{\Delta ILC}$ mice exhibited intestinal inflammation characterized by crypt elongation, loss of normal architecture and significantly decreased colon length (Fig. 3f, g). Intestinal inflammation in MHCII$^{\Delta ILC}$ mice could be prevented by continuous administration of antibiotics (Fig. 3e–g), demonstrating a critical role for commensal bacteria in the development of disease. Moreover, MHCII$^{\Delta ILC}$ mice exhibited significantly elevated frequencies of IFN-γ$^+$, IL-17A$^+$ and TNF-α$^+$ CD4$^+$ T cells in the colon (Fig. 3h), including CD4$^+$ T cells that co-produced IFN-γ and IL-17A (Fig. 3i), which was associated with significant recruitment of neutrophils into the colonic lamina propria (Fig. 3j). Further phenotypic and functional analyses suggested that the increased pro-inflammatory CD4$^+$ T-cell responses and intestinal inflammation that developed in the absence of ILC-intrinsic MHCII were not the result of impaired regulatory T-cell function or regulatory cytokine production (Supplementary Fig. 10), altered thymic selection (Supplementary Fig. 11) or commensal microflora dysbiosis (Supplementary Fig. 12). Collectively, these data suggest that ILCs directly limit commensal bacteria-responsive CD4$^+$ T cells through MHCII-dependent interactions.

To determine whether dysregulated CD4$^+$ T-cell responses to commensal bacteria directly promoted intestinal inflammation observed in MHCII$^{\Delta ILC}$ mice, sort-purified CD4$^+$ T cells from control *H2-Ab1*^{fl} mice or MHCII$^{\Delta ILC}$ mice were transferred into *Rag1*$^{-/-}$ mice. In comparison to *Rag1*$^{-/-}$ recipients receiving CD4$^+$ T cells from control mice, *Rag1*$^{-/-}$ recipients receiving CD4$^+$ T cells from MHCII$^{\Delta ILC}$ mice exhibited rapid and substantial weight loss (Fig. 4a), colonic shortening, macroscopic intestinal thickening and severe intestinal inflammation characterized by crypt elongation, loss of normal architecture and inflammatory cell infiltrates (Fig. 4b–d). Critically, oral administration of antibiotics to MHCII$^{\Delta ILC}$ donor mice before CD4$^+$ T-cell isolation abrogated the ability of CD4$^+$ T cells to elicit wasting disease and intestinal inflammation in naive *Rag1*$^{-/-}$ recipients (Fig. 4a–d). Similarly, germ-free *Rag1*$^{-/-}$ recipients of CD4$^+$ T cells from MHCII$^{\Delta ILC}$ mice did not develop wasting disease or intestinal inflammation in comparison to conventional *Rag1*$^{-/-}$ recipients (Fig. 4e–g). Therefore, in the absence of RORγt$^+$ ILC-intrinsic MHCII, commensal bacteria are required for both the development of pathological CD4$^+$ T-cell responses and for the onset of wasting disease and intestinal inflammation.

Collectively, these data identify a previously unrecognized role for RORγt$^+$ ILCs in maintaining intestinal homeostasis by limiting pathological CD4$^+$ T-cell responses to commensal bacteria through MHCII-dependent interactions (Supplementary Fig. 13). MHCII expression was found to be restricted to a subset of CCR6$^+$RORγt$^+$ ILCs that lack T-bet and IFN-γ expression and thus, are phenotypically and functionally distinct from RORγt$^+$ ILC populations that have been associated with promoting intestinal inflammation in murine models and inflammatory bowel disease patients[2,11]. Rather, MHCII$^+$RORγt$^+$

ILCs are more similar to IL-22-producing ILC populations previously shown to promote tissue protection[3,5,8,12–14]. In a developmental context, it is remarkable that RORγt$^+$ ILCs are the first cells of the immune system to colonize the neonatal intestine and gut-associated lymphoid tissues[15,25,26]. Therefore, ILCs may have a critical role not only in promoting lymphoid organogenesis and cytokine-mediated epithelial cell barrier integrity but also in regulating adaptive immune cell responses to newly colonizing commensal bacteria. It may also be advantageous for ILCs to modulate commensal bacteria-specific T cells as, in contrast to professional antigen presenting cells, murine ILCs lack expression of TLRs, conventional co-stimulatory molecules and do not produce cytokines that regulate T-cell differentiation[1,8,10,25]. The demonstration that ILCs regulate adaptive immune cell responses to commensal bacteria through a MHCII-dependent mechanism may be of importance in understanding the pathogenesis of numerous chronic human diseases associated with inflammatory host immune responses to commensal bacteria.

METHODS SUMMARY

MHCII$^{\Delta ILC}$ mice were generated by crossing *H2-Ab1*^{fl} mice with *Rorc*^{cre} mice. *H2-Ab1*^{fl} littermates were used as controls for all experiments. Intestinal lamina propria and lymphoid tissue cell suspensions were isolated as previously described[3,7]. In some experiments mice were treated with antibiotics as previously described[3]. Commensal bacteria-derived antigen-specific enzyme-linked immunosorbent assays (ELISAs) were performed by coating 96-well plates with 5 μg ml^{-1} crude commensal bacteria antigen, prepared by isolation of faecal contents and sequential homogenization and sonication. ILCs were identified by exclusion of lineage-positive cells by staining with fluorochrome-conjugated antibodies against CD3, CD5, CD8α, CD11c, NK1.1 and B220 and by positive expression of CD45 and CD90.2 or by expression of GFP in RORγt–eGFP mice. Sort-purified ILCs and dendritic cells were assessed for antigen processing and presentation capacity by culturing cells with 10 μg ml^{-1} DQ-OVA for 3 h or by culturing for 3 h with 50 μg ml^{-1} GFP-labelled E-alpha protein, followed by subsequent staining with an antibody recognizing Eα_{52-68} peptide bound to I–A^b. Lymphocyte populations (≥95%) were sort-purified using a BD FACS Aria II. CD4$^+$ T cells were adoptively transferred to conventional or germ-free *Rag1*$^{-/-}$ recipients via intravenous injection. Data are represented as the mean ± s.e.m. and statistical significance was determined by the Student's *t*-test.

Full Methods and any associated references are available in the online version of the paper.

Received 4 November 2012; accepted 2 May 2013.
Published online 22 May 2013.

1. Spits, H. & Cupedo, T. Innate lymphoid cells: emerging insights in development, lineage relationships, and function. *Annu. Rev. Immunol.* **30,** 647–675 (2012).
2. Buonocore, S. *et al.* Innate lymphoid cells drive interleukin-23-dependent innate intestinal pathology. *Nature* **464,** 1371–1375 (2010).
3. Sonnenberg, G. F. *et al.* Innate lymphoid cells promote anatomical containment of lymphoid-resident commensal bacteria. *Science* **336,** 1321–1325 (2012).
4. Cella, M. *et al.* A human natural killer cell subset provides an innate source of IL-22 for mucosal immunity. *Nature* **457,** 722–725 (2009).
5. Sawa, S. *et al.* RORγt$^+$ innate lymphoid cells regulate intestinal homeostasis by integrating negative signals from the symbiotic microbiota. *Nature Immunol.* **12,** 320–326 (2011).
6. Lochner, M. *et al.* Microbiota-induced tertiary lymphoid tissues aggravate inflammatory disease in the absence of RORγt and LTi cells. *J. Exp. Med.* **208,** 125–134 (2011).
7. Sonnenberg, G. F., Monticelli, L. A., Elloso, M. M., Fouser, L. A. & Artis, D. CD4(+) lymphoid tissue-inducer cells promote innate immunity in the gut. *Immunity* **34,** 122–134 (2011).
8. Sonnenberg, G. F. & Artis, D. Innate lymphoid cell interactions with microbiota: implications for intestinal health and disease. *Immunity* **37,** 601–610 (2012).
9. Spits, H. *et al.* Innate lymphoid cells—a proposal for uniform nomenclature. *Nature Rev. Immunol.* **13,** 145–149 (2013).
10. Walker, J. A., Barlow, J. L. & McKenzie, A. N. Innate lymphoid cells—how did we miss them? *Nature Rev. Immunol.* **13,** 75–87 (2013).
11. Geremia, A. *et al.* IL-23-responsive innate lymphoid cells are increased in inflammatory bowel disease. *J. Exp. Med.* **208,** 1127–1133 (2011).
12. Takayama, T. *et al.* Imbalance of NKp44$^+$NKp46$^-$ and NKp44$^-$NKp46$^+$ natural killer cells in the intestinal mucosa of patients with Crohn's disease. *Gastroenterology* **139,** 882–892 (2010).
13. Ciccia, F. *et al.* Interleukin-22 and IL-22-producing NKp44$^+$ NK cells in the subclinical gut inflammation of patients with ankylosing spondylitis. *Arthritis Rheum.* **64,** 1869–1878 (2011).
14. Sonnenberg, G. F., Fouser, L. A. & Artis, D. Border patrol: regulation of immunity, inflammation and tissue homeostasis at barrier surfaces by IL-22. *Nature Immunol.* **12,** 383–390 (2011).
15. Eberl, G. & Littman, D. R. Thymic origin of intestinal αβ T cells revealed by fate mapping of RORγt$^+$ cells. *Science* **305,** 248–251 (2004).
16. Sawa, S. *et al.* Lineage relationship analysis of RORγt$^+$ innate lymphoid cells. *Science* **330,** 665–669 (2010).
17. Monticelli, L. A. *et al.* Innate lymphoid cells promote lung-tissue homeostasis after infection with influenza virus. *Nature Immunol.* **12,** 1045–1054 (2011).
18. Reynders, A. *et al.* Identity, regulation and *in vivo* function of gut NKp46$^+$RORγt$^+$ and NKp46$^+$RORγt$^-$ lymphoid cells. *EMBO J.* **30,** 2934–2947 (2011).
19. Yosef, N. *et al.* Dynamic regulatory network controlling T17 cell differentiation. *Nature* **496,** 461–468 (2013).
20. Bezman, N. A. *et al.* Molecular definition of the identity and activation of natural killer cells. *Nature Immunol.* **13,** 1000–1009 (2012).
21. Klose, C. S. *et al.* A T-bet gradient controls the fate and function of CCR6-RORγt$^+$ innate lymphoid cells. *Nature* **494,** 261–265 (2013).
22. Schwartz, R. H. T cell anergy. *Annu. Rev. Immunol.* **21,** 305–334 (2003).
23. Cong, Y., Feng, T., Fujihashi, K., Schoeb, T. R. & Elson, C. O. A dominant, coordinated T regulatory cell-IgA response to the intestinal microbiota. *Proc. Natl Acad. Sci. USA* **106,** 19256–19261 (2009).
24. Benoist, C. & Mathis, D. Regulation of major histocompatibility complex class-II genes: X, Y and other letters of the alphabet. *Annu. Rev. Immunol.* **8,** 681–715 (1990).
25. Mebius, R. E., Rennert, P. & Weissman, I. L. Developing lymph nodes collect CD4+CD3- LTbeta+ cells that can differentiate to APC, NK cells, and follicular cells but not T or B cells. *Immunity* **7,** 493–504 (1997).
26. Eberl, G. *et al.* An essential function for the nuclear receptor RORγt in the generation of fetal lymphoid tissue inducer cells. *Nature Immunol.* **5,** 64–73 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Sonnenberg and Artis laboratories for discussions and critical reading of the manuscript. We also thank H. L. Ma, L. A. Fouser, S. Olland, R. Zollner, K. Lam and A. Root at Pfizer for critical discussions, valuable advice and the preparation of IL-22 antibodies; M. M. Elloso at Janssen Research and Development for critical discussions, valuable advice and the preparation of IL-17 and IL-23 antibodies; and M.S. Marks for providing the E-alpha protein and Y-Ae antibody. The research is supported by the National Institutes of Health (AI061570, AI087990, AI074878, AI095776, AI102942, AI095466, AI095608 and AI097333 to D.A.; T32-AI055428 to L.A.M.; DK071176 to C.O.E.; and DP5OD012116 to G.F.S.), the Crohn's and Colitis Foundation of America (to D.A.) and the Burroughs Wellcome Fund Investigator in Pathogenesis of Infectious Disease Award (to D.A.). We also thank the Matthew J. Ryan Veterinary Hospital Pathology Lab, the National Institute of Diabetes and Digestive and Kidney Disease Center for the Molecular Studies in Digestive and Liver Disease Molecular Pathology and Imaging Core (P30DK50306), the Penn Microarray Facility and the Abramson Cancer Center Flow Cytometry and Cell Sorting Resource Laboratory (partially supported by NCI Comprehensive Cancer Center Support Grant (2-P30 CA016520)) for technical advice and support. Human tissue samples were provided by the Cooperative Human Tissue Network, which is funded by the National Cancer Institute.

Author Contributions M.R.H., L.A.M., T.C.F., D.A. and G.F.S. designed and performed the research. C.G.K.Z. performed analyses of microarray data. A.C., J.M.A. and E.J.W. performed the microarray of naive CD4$^+$ T cells. S.G., R.S. and F.D.B. provided advice, performed and analysed 454 pyrosequencing of intestinal commensal bacteria. A.R.M. assisted with reagents and advice for antigen presentation assays. H.-L.M., P.A.K., C.O.E. and G.E. provided essential mouse strains, valuable advice and technical expertise for these studies. M.R.H., L.A.M., T.C.F., D.A. and G.F.S. analysed the data. M.R.H., D.A. and G.F.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.F.S. (gfield@mail.med.upenn.edu).

METHODS

Mice, antibiotics and use of monoclonal antibodies *in vivo*. C57BL/6 mice, C57BL/6 *Rag1*$^{-/-}$, C57BL/6 CD90.1 and C57BL/6 *Rorc*$^{gfp/gfp}$ mice were purchased from the Jackson Laboratory, bred and maintained at the University of Pennsylvania. C57BL/6 *Il17a*$^{-/-}$ mice were provided by Y. Iwakura (University of Tokyo), C57BL/6 *Il22*$^{-/-}$ mice were provided by Pfizer, *Il23a*$^{-/-}$ mice were provided by Janssen Research & Development LLC, *H2-Ab1*^{fl} mice were provided by P. A. Koni, tissues from CBir1 TCR transgenic mice were provided by C. O. Elson, and *Rorc*^{cre} mice and *Rorc(γt)-Gfp*^{TG} were provided by G. Eberl. All mice were maintained in specific pathogen-free facilities at the University of Pennsylvania. Germ-free C57BL/6 and C57BL/6 *Rag1*$^{-/-}$ mice were provided by the University of Pennsylvania Gnotobiotic Mouse Facility. CD90-disparate *Rag*1$^{-/-}$ chimaeras were constructed as previously described[3,7]. All protocols were approved by the University of Pennsylvania Institutional Animal Care and Use Committee (IACUC), and all experiments were performed according to the guidelines of the University of Pennsylvania IACUC. A previously described cocktail of antibiotics was continuously administered via drinking water for defined periods of time[3,27]. Anti-CD90.2 monoclonal antibody (30H12) was purchased from BioXCell. Anti-IL-22 monoclonal antibodies, IL22-01 (neutralizing) and IL22-02 (mouse cytokine detection) were developed by Pfizer. Anti-IL-17A monoclonal antibody (CNTO 8096) and anti-IL-23p19 monoclonal antibody (CNTO 6163) were developed by Janssen Research & Development, LLC. Anti-IL-17RA monoclonal antibody was developed by Amgen Inc. Neutralizing or depleting monoclonal antibodies were administered intraperitoneally every 3 days at a dose of 250 μg per mouse starting on day 0 and ending on day 14.

Murine tissue isolation and flow cytometry. Spleens, lymph nodes and Peyer's patches were harvested, and single-cell suspensions were prepared at necropsy as previously described[3,7]. For intestinal lamina propria lymphocyte preparations, intestines were isolated, attached fat removed and tissues cut open longitudinally. Luminal contents were removed by shaking in cold PBS. Epithelial cells and intraepithelial lymphocytes were removed by shaking tissue in stripping buffer (1 mM EDTA, 1 mM DTT and 5% FCS) for 30 min at 37 °C. The lamina propria layer was isolated by digesting the remaining tissue in 0.5 mg ml^{-1} collagenase D (Roche) and 20 μg ml^{-1} DNase I (Sigma-Aldrich) for 30 min at 37 °C.

For flow cytometric analyses, cells were stained with antibodies to the following markers: anti-NK1.1 (clone PK136, eBioscience), anti-CD3 (clone 145-2C11, eBioscience), anti-CD5 (clone 53-7.3, eBioscience), anti-CD90.2 (clone 30-H12, BioLegend), anti-CD127 (clone A7R34, eBioscience), anti-CD11c (clone N418, eBioscience), anti-F4/80 (clone BM8, eBioscience), anti-CD4 (clone GK1.5, Abcam), anti-CD8 (clone 53-6.7, eBioscience), anti-B220 (clone RA3-6B2, eBioscience), anti-CD25 (clone eBio3C7, eBioscience), anti-MHCII (clone M5/114.15.2, eBioscience), anti-CD44 (clone IM7, eBioscience), anti-CD62L (clone MEL-14, eBioscience), anti-CD45 (clone 30-F11, eBioscience), anti-NKp46 (clone 29A1.4, eBioscience), anti-CD11b (clone M1/70, eBioscience), anti-CD117 (c-kit) (clone 2B8, eBioscience), anti-Sca-1 (clone D7, eBioscience), anti-CD40 (clone 1C10, eBioscience), anti-CD80 (clone 16-10A1, eBioscience), anti-CD86 (clone GL1, BD Biosciences), anti-Ly6G (clone 1A8, BioLegend) and anti-CCR6 (clone 29-2L17, BioLegend). For intracellular staining, cells were fixed and permeabilized using a commercially available kit (eBioscience) and stained with anti-RORγt (clone B2D, eBioscience), anti-FoxP3 (clone FJK016s, eBioscience), anti-T-bet (clone eBio-4B10, eBioscience), anti-GATA-3 (clone TWAJ, eBioscience) or anti-Ki-67 (clone B56, BD Biosciences). For cytokine production, cells were stimulated *ex vivo* by incubation for 4 h with 50 ng ml^{-1} PMA, 750 ng ml^{-1} ionomycin, 10 μg ml^{-1} brefeldin A (all obtained from Sigma-Aldrich) or 50 ng ml^{-1} rIL-23 (eBioscience) and 10 μg ml^{-1} brefeldin A. Cells were fixed and permeabilized as indicated above and stained with IL22-02 (Pfizer) conjugated to Alexa Fluor 647 or Alexa Fluor 488 according to the manufacturer's instructions (Molecular Probes), anti-IL-17A (clone eBioTC11-18H10.1, eBioscience), anti-IFN-γ (clone XMG1.2, eBioscience) and anti-TNF-α (clone MP6-XT22, eBioscience). Dead cells were excluded from analysis using a violet viability stain (Invitrogen). T-cell Vβ chain usage was assessed using a commercial mouse Vβ TCR screening panel (BD Biosciences). Flow cytometry data collection was performed on a LSR II (BD Biosciences) and cell sorting performed on an Aria II (BD Biosciences). Data were analysed using FlowJo software (Tree Star Inc.).

Human intestinal samples and flow cytometry. Human intestinal tissues from the ileum were obtained from the Cooperative Human Tissue Network. Single cell suspensions from intestinal tissues were obtained by cutting tissues into small pieces and incubating for 1–2 h at 37 °C with shaking in stripping buffer (1 mM EDTA, 1 mM DTT and 5% FCS) to remove the epithelial layer. Supernatants were then discarded, and the lamina propria fraction was obtained by incubating the remaining tissue for 1–2 h at 37 °C with shaking in collagenase solution. Remaining tissues were then mechanically dissociated, filtered through a wire mesh tissue sieve, and lymphocytes were subsequently separated by Ficoll gradient.

For flow cytometry, cells were stained with antibodies to the following markers: anti-CD3 (clone UCHT1, eBioscience), anti-CD56 (clone CMSSB, eBioscience), anti-CD19 (clone 2H7, eBioscience), anti-HLA-DR (clone LN3, eBioscience), anti-CD127 (clone A019D5, Biolegend). For intracellular staining, cells were fixed and permeabilized using a commercially available kit (eBioscience) and stained with anti-RORγt (clone AFKJS-9, eBioscience) and anti-IL-22 (clone 22URTI, eBioscience). Dead cells were excluded from analysis using a viability stain (Invitrogen). Flow cytometry data was collected using a LSR II (BD Biosciences). Data were analysed using FlowJo software (Tree Star Inc.).

Histological sections. Tissue samples from the intestines of mice were fixed with 4% paraformaldehyde, embedded in paraffin, and 5 μm sections were stained with haematoxylin and eosin.

Microarray and DAVID pathway analysis. Microarray gene expression profiling and data normalization for group 2 ILCs (sorted Lineage$^-$, CD90.2$^+$, CD25$^+$ ILCs from the lungs of naive C57BL/6 mice), group 3 ILCs (Lineage$^-$, CD90.2$^+$, CD4$^+$ ILCs from the spleens of naive C57BL/6 mice) and naive splenic CD4$^+$ T cells were performed as previously described (GEO accession numbers; GSE46468 and GSE30437), data were RMA-normalized and SAM analysis was performed[17]. Additional microarray gene expression profiling were obtained from GEO for previously published studies of *in vitro* polarized T_H17 cells (GSM1074979, GSM1075001 and GSM1075002)[19], splenic NK cells (GSM538315, GSM538316 and GSM538317)[20] and subsets of RORγ$^+$ ILCs isolated from naive murine small intestine (GSM739586, GSM739591, GSM739588, GSM739593, GSM739589 and GSM739594)[18]. GEO data sets were batch-corrected to existing microarray gene expression profiles using ComBat[28]. Differentially expressed genes in the transcriptional profiles of specified groups were uploaded to the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/)[29] and analysed as previously described[17] using the Fisher's exact test to identify significantly enriched Gene Ontology (GO, http://www.geneontology.org) terms[29]. Heat maps displaying key genes were generated using Mayday software[30].

Antigen processing and presentation analyses. 2×10^3 sort-purified CD11c$^+$MHCII$^+$ dendritic cells and lineage$^-$CD127$^+$c-kit$^+$MHCII$^+$ ILCs were cultured with 10 μg ml^{-1} DQ-OVA (Molecular Probes) at 4 °C or 37 °C for 3 h, extensively washed and fluorescence assessed via flow cytometry. Alternatively, antigen processing and presentation were assessed using a previously defined self-antigen E-alpha[31]. Briefly, 2×10^3 sort-purified dendritic cells and ILCs were pulsed for 3 h in complete media with 50 μg ml^{-1} GFP-labelled E-alpha protein, extensively washed and stained with a biotin-conjugated antibody recognizing $E\alpha_{52-68}$ peptide bound to I-A^b (clone Yae, eBioscience) followed by a streptavidin APC (eBioscience) and uptake of antigen and peptide-MHCII complex presentation assessed via flow cytometry. In some assays 2×10^3 sort-purified dendritic cells or ILCs were pulsed with 50 μg ovalbumin for 2 h before incubation with 2×10^4 sort-purified CFSE-labelled OT-II T cells. Cell co-cultures were incubated for 72 h before analysis via flow cytometry.

T-cell adoptive transfer. 2×10^6 CD4$^+$CD3$^+$ T cells were sorted from the spleen and mesenteric lymph node of control or experimental mice to a purity >97% and transferred intravenously to naive C57BL/6 *Rag1*$^{-/-}$ recipient mice (GF or CNV). Weights of recipient mice were monitored through the progression of the experiment.

CBir1-specific T-cell transfers. 1×10^6 CBir1 TCR transgenic T cells were transferred into congenically marked hosts with or without co-transfer of 8×10^3 sort-purified ILCs (lineage$^-$ CD127$^+$ c-kit$^+$) pulsed with 1 μg ml^{-1} $CBir1_{456-475}$ peptide. Twenty-four hours later mice were administered 50 μg $CBir1_{456-475}$ peptide intraperitoneally and 72 h later mice were euthanized and the presence of CBir1 TCR transgenic T cells was quantified in the spleen. IFN-γ production was quantified by culturing 1×10^6 splenocytes in the presence of 1 μg ml^{-1} $CBir1_{456-475}$ peptide for 48 h.

Regulatory T-cell suppression assay. CD11c$^+$ dendritic cells, naive CD4$^+$CD25$^-$ CD45RB^{hi} T effector (T_{eff}) cells and CD4$^+$CD25$^+$CD45RB$^-$ regulatory T cells (T_{reg}) cells were sort-purified from the spleen and mesenteric lymph node of MHCII$^{\Delta ILC}$ mice and littermate controls. Sort-purified T_{reg} cells were found to be at least 98% FoxP3$^+$. Dendritic cells were plated at 5×10^3 per well in the presence or absence of 1 μg ml^{-1} soluble purified anti-CD3 (clone 145-2C11, BD Biosciences). T_{eff} cells were CFSE labelled and added to wells containing dendritic cells at 2.5×10^4 alone, or with T_{reg} at a ratio of 1:0, 1:2, 1:4, 1:8 and 1:16. After 3 days culture at 37 °C/5% CO_2, cell-culture supernatants were harvested and T_{eff} proliferation was measured by CFSE dilution via flow cytometry. T_{reg} suppression was calculated by gating on T effector cells and quantifying the percentage of CFSE-dim in comparison to cells cultured in the absence of T_{reg} cells.

Quantitative real-time PCR. RNA was isolated from whole colon tissue that was homogenized and snap frozen in Trizol reagent (Invitrogen). RNA was isolated as per the manufacturer's instructions and cDNA generated using Superscript reverse transcription (Invitrogen). Real-time PCR was performed on cDNA using

SYBR green chemistry (Applied Biosystems) using commercially available primer sets (QIAGEN). Reactions were run on a real-time PCR system (ABI7500; Applied Biosystems). Samples were normalized to β-actin and displayed as a fold change as compared to control mice.

Commensal bacteria-specific ELISA. Colonic faecal contents were homogenized and briefly centrifuged at 1,000 r.p.m. to remove large aggregates, and the resulting supernatant was washed with sterile PBS twice by centrifuging for 1 min at 8,000 r.p.m. On the last wash, bacteria were re-suspended in 2 ml ice-cold PBS and sonicated on ice. Samples were then centrifuged at 20,000*g* for 10 min, and supernatants recovered for a crude commensal bacteria antigen preparation. For measurement of serum antibodies by ELISA, 5 μg ml^{-1} commensal bacteria antigen was coated on 96-well plates, and sera were incubated in doubling dilutions. Antigen-specific IgG was detected using an anti-mouse IgG-HRP antibody (BD Biosciences). Plates were developed with TMB peroxidase substrate (KPL), and optical densities measured using a plate spectrophotometer.

Microbiota transfer to germ-free mice. The caeca of MHCII$^{\Delta ILC}$ mice and littermate controls were opened under aseptic conditions and caecal contents re-suspended in sterile PBS. Germ-free C57BL/6 mice were then orally gavaged with 200 μl of caecal content suspension and subsequently monitored over the course of 6 weeks for signs of disease and rectal prolapse before death.

Pyrosequencing. DNA from luminal contents from the large intestine of mice was obtained using the QIAamp DNA Stool mini kit (Qiagen). DNA samples were amplified using the V1-V2 region primers targeting bacterial 16S genes and sequenced using 454/Roche Titanium technology. Sequence analysis was carried out using the QIIME pipeline[32] for co-housed cohorts of *H2-Ab1*^{fl} and MHCII$^{\Delta ILC}$ mice.

Statistical analysis. Results represent the mean ± s.e.m. Statistical significance was determined by the Student's *t*-test ($*P < 0.05$; $**P < 0.01$; $***P < 0.001$).

27. Abt, M. C. *et al.* Commensal bacteria calibrate the activation threshold of innate antiviral immunity. *Immunity* **37,** 158–170 (2012).
28. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2007).
29. Huang, D. W. *et al.* Extracting biological meaning from large gene lists with DAVID. *Curr. Protoc. Bioinform.* Ch. 13, Unit 13 11 (2009).
30. Battke, F., Symons, S. & Nieselt, K. Mayday–integrative analytics for expression data. *BMC Bioinformatics* **11,** 121 (2010).
31. Murphy, D. B. *et al.* A novel MHC class II epitope expressed in thymic medulla but not cortex. *Nature* **338,** 765–768 (1989).
32. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).

LETTER

doi:10.1038/nature12166

Control of angiogenesis by AIBP–mediated cholesterol efflux

Longhou Fang[1], Soo-Ho Choi[1], Ji Sun Baek[1], Chao Liu[1], Felicidad Almazan[1], Florian Ulrich[2], Philipp Wiesner[1], Adam Taleb[1], Elena Deer[1], Jennifer Pattison[1], Jesús Torres-Vázquez[2], Andrew C. Li[1]‡ & Yury I. Miller[1]

Cholesterol is a structural component of the cell and is indispensable for normal cellular function, although its excess often leads to abnormal proliferation, migration, inflammatory responses and/or cell death. To prevent cholesterol overload, ATP-binding cassette (ABC) transporters mediate cholesterol efflux from the cells to apolipoprotein A-I (apoA-I) and the apoA-I-containing high-density lipoprotein (HDL)[1–3]. Maintaining efficient cholesterol efflux is essential for normal cellular function[4–6]. However, the role of cholesterol efflux in angiogenesis and the identity of its local regulators are poorly understood. Here we show that apoA-I binding protein (AIBP) accelerates cholesterol efflux from endothelial cells to HDL and thereby regulates angiogenesis. AIBP- and HDL-mediated cholesterol depletion reduces lipid rafts, interferes with VEGFR2 (also known as KDR) dimerization and signalling and inhibits vascular endothelial growth factor-induced angiogenesis *in vitro* and mouse aortic neovascularization *ex vivo*. Notably, Aibp, a zebrafish homologue of human AIBP, regulates the membrane lipid order in embryonic zebrafish vasculature and functions as a non-cell-autonomous regulator of angiogenesis. *aibp* knockdown results in dysregulated sprouting/branching angiogenesis, whereas forced Aibp expression inhibits angiogenesis. Dysregulated angiogenesis is phenocopied in Abca1 (also known as Abca1a) Abcg1-deficient embryos, and cholesterol levels are increased in Aibp-deficient and Abca1 Abcg1-deficient embryos. Our findings demonstrate that secreted AIBP positively regulates cholesterol efflux from endothelial cells and that effective cholesterol efflux is critical for proper angiogenesis.

AIBP is a secreted protein that was discovered in a screen of proteins that physically associate with apoA-I[7]. Human *APOA1BP* mRNA encoding the AIBP protein is ubiquitously expressed[7]. Although AIBP binding to apoA-I implies that AIBP may modulate HDL function[7,8], its role in cholesterol efflux has not been experimentally tested.

First, we investigated whether human AIBP had any effect on cholesterol removal from human umbilical vein endothelial cells (HUVECs), in which ABCG1 is a key transporter responsible for cholesterol efflux

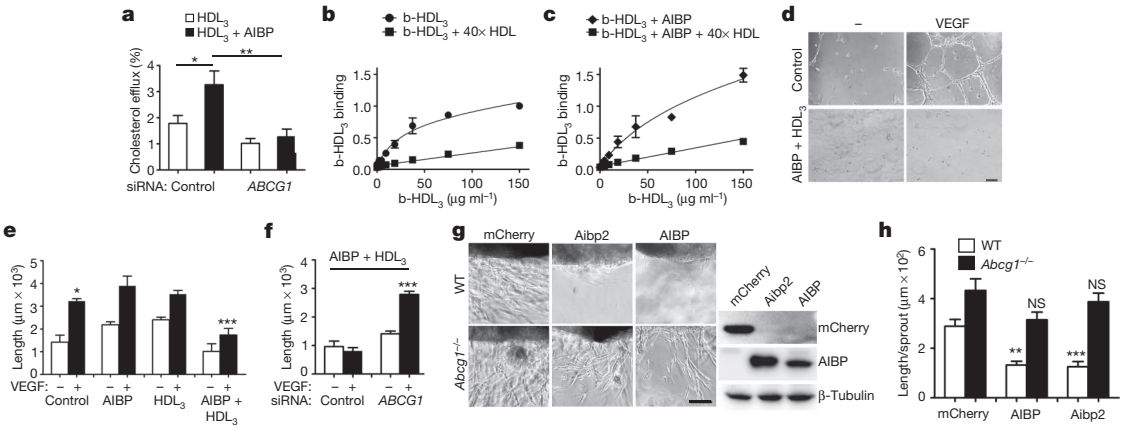


Figure 1 | Role of AIBP in cholesterol efflux from endothelial cells and *in vitro* angiogenesis. a, Human-AIBP-mediated cholesterol efflux and effect of *ABCG1* knockdown. HUVECs were transfected with control or *ABCG1* siRNA, preloaded with ^{3}H-cholesterol and incubated for 1 h with 50 μg ml^{-1} HDL_3 in the presence or absence of 0.2 μg ml^{-1} AIBP. Efflux was measured as the ^{3}H counts in the medium divided by the sum of ^{3}H counts in the medium and the cells. Mean ± s.e.; $n = 6$. **b**, **c**, Effect of human AIBP on HDL_3 binding to HUVECs. HUVECs were incubated on ice with the indicated concentration of biotinylated HDL_3 (b-HDL_3), in the presence or absence of AIBP (at a 0.1:50 (w/w) AIBP:HDL_3 ratio) and 40× excess of unlabelled HDL. Each data point is mean ± s.e. from 3 to 7 independent experiments. The binding parameters for b-HDL_3–HUVEC binding were calculated as $B_{max} = 0.8 \pm 0.1$ and $K_d = (0.33 \pm 0.10) \times 10^{-6}$ M in absence of AIBP (panel **b**; $R^2 = 0.92$, Sy.x = 0.1) and $B_{max} = 1.5 \pm 0.4$ and $K_d = (1.03 \pm 0.5) \times 10^{-6}$ M in the presence of AIBP (panel **c**; $R^2 = 0.94$, Sy.x = 0.1). The differences in B_{max} and K_d values were statistically significant ($P < 0.01$ and $P < 0.05$, respectively). **d**, Effect of human AIBP and HDL_3 on endothelial cell tube formation. HUVECs were pre-incubated with or without 50 μg ml^{-1} HDL_3 + 0.1 μg ml^{-1} AIBP for 4 h. Cells were then seeded on Matrigel, in the presence or absence of 20 ng ml^{-1} VEGF, and imaged following a 12-h incubation. Scale bar, 100 μm. **e**, The length of endothelial cell tubes in the experiment illustrated in **d** and Supplementary Fig. 4. Mean ± s.e.; $n = 5$. **f**, Requirement for ABCG1 in human AIBP inhibition of angiogenesis. HUVECs were transfected with control or *ABCG1* siRNA and assayed as in **d**. Mean ± s.e.; $n = 6$. **g**, Mouse aortic ring angiogenesis assay. Aortic rings from C57BL/6 wild-type (WT) and *Abcg1*$^{-/-}$ mice were embedded in Matrigel. Human HEK293 cells transiently expressing mCherry, zebrafish Aibp2 or human AIBP were inserted approximately 0.5 mm away from the aortic ring, and the plates were incubated with 10 ng ml^{-1} VEGF for 7 days. Images show the edge of the aortic rings facing the HEK293 cell clusters. Immunoblots show expression of AIBP and Aibp2 (both detected with a Flag tag antibody) and mCherry in HEK293 cells. **h**, The length of aortic ring sprouts. Mean ± s.e.; $n = 10$. NS, not significant. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

[1]Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [2]Helen L. and Martin S. Kimmel Center for Biology and Medicine, Skirball Institute of Biomolecular Medicine, New York University Langone Medical Center, 540 First Avenue, New York, New York 10016, USA.
‡Deceased.

to HDL[9,10]. In the presence of AIBP, cholesterol efflux from HUVECs to HDL_3 (a subfraction of HDL, which is an efficient cholesterol acceptor) was increased twofold, an effect completely abrogated by ABCG1 deficiency (Fig. 1a and Supplementary Fig. 2). AIBP did not promote cholesterol efflux in the absence of HDL_3 (Supplementary Fig. 2b), but the binding of AIBP to HUVECs (Supplementary Fig. 3) increased the overall HUVEC capacity to bind HDL_3 ($B_{max} = 1.5$ versus 0.8) and the constant of HDL_3 dissociation from HUVECs ($K_d = 1.0 \times 10^{-6}$ M versus 0.33×10^{-6} M; Fig. 1b, c), thereby creating conditions that would facilitate HDL_3-mediated cholesterol efflux.

To investigate the role of AIBP- and HDL-mediated cholesterol efflux in angiogenesis, we incubated HUVECs with AIBP and/or HDL_3 and then stimulated cells with vascular endothelial growth factor (VEGF). AIBP and HDL_3 added separately did not affect endothelial cell tube formation, but together they significantly reduced angiogenesis (Fig. 1d, e and Supplementary Fig. 4). Cholesterol depletion by methyl-β-cyclodextrin (MβCD)[11] also inhibited angiogenesis, whereas cholesterol-loaded MβCD, which delivers cholesterol to the cell, promoted angiogenesis (Supplementary Fig. 5). If the AIBP- and HDL_3-mediated inhibition of angiogenesis is the consequence of accelerated cholesterol efflux, then this effect should depend on the presence of the cholesterol transporter ABCG1. Indeed, knockdown of *ABCG1* in HUVECs rescued VEGF-induced angiogenesis from the AIBP- and HDL_3-mediated inhibition (Fig. 1f). Further, we tested both human AIBP and a zebrafish AIBP homologue, here termed Aibp2 (discussed in more detail below), in an *ex vivo* aortic ring angiogenesis assay. A cluster of HEK293 cells producing either AIBP, Aibp2 or mCherry (used as a negative control) was placed 0.5 mm from the edge of a mouse aortic ring, and VEGF was added to stimulate angiogenesis. Both AIBP and Aibp2, but not mCherry, significantly reduced neovascularization of aortic rings isolated from a wild-type mouse (Fig. 1g, h). Aortic rings from an $Abcg1^{-/-}$ mouse responded to VEGF with a more vigorous angiogenesis, which was not significantly reduced by AIBP or Aibp2. These results support the hypothesis that cholesterol efflux is necessary for the AIBP-mediated inhibition of angiogenesis.

HDL-mediated depletion of cholesterol from plasma membrane disrupts cholesterol- and sphingomyelin-rich membrane microdomains[12,13], often designated as lipid rafts, and affects membrane receptor signalling[11]. We found that human AIBP and HDL_3 reduced the lipid raft content in HUVECs and disrupted cell-surface co-localization of caveolin 1 (CAV1) and VEGFR2 (Fig. 2a–c and Supplementary Fig. 6). The AIBP/HDL_3 treatment, similarly to the treatment with MβCD, decreased VEGFR2 and CAV1 localization to the lipid raft fraction isolated from cell lysates (Fig. 2d and Supplementary Fig. 7). Many studies suggest that VEGFR2 localization to lipid rafts facilitates VEGFR2 dimerization and endocytosis[14–17], the steps required for VEGF-mediated signalling[18]. In our experiments, AIBP/HDL_3 treatment reduced VEGF-induced VEGFR2 dimerization and endocytosis as well as phosphorylation of VEGFR2, AKT, FAK (also known as PTK2), SRC, and to a lesser degree of ERK1 (also known as MAPK3) and ERK2 (also known as MAPK1) (Fig. 2e–g and Supplementary Figs 8–10). Notably, subsequent addition of cholesterol partially

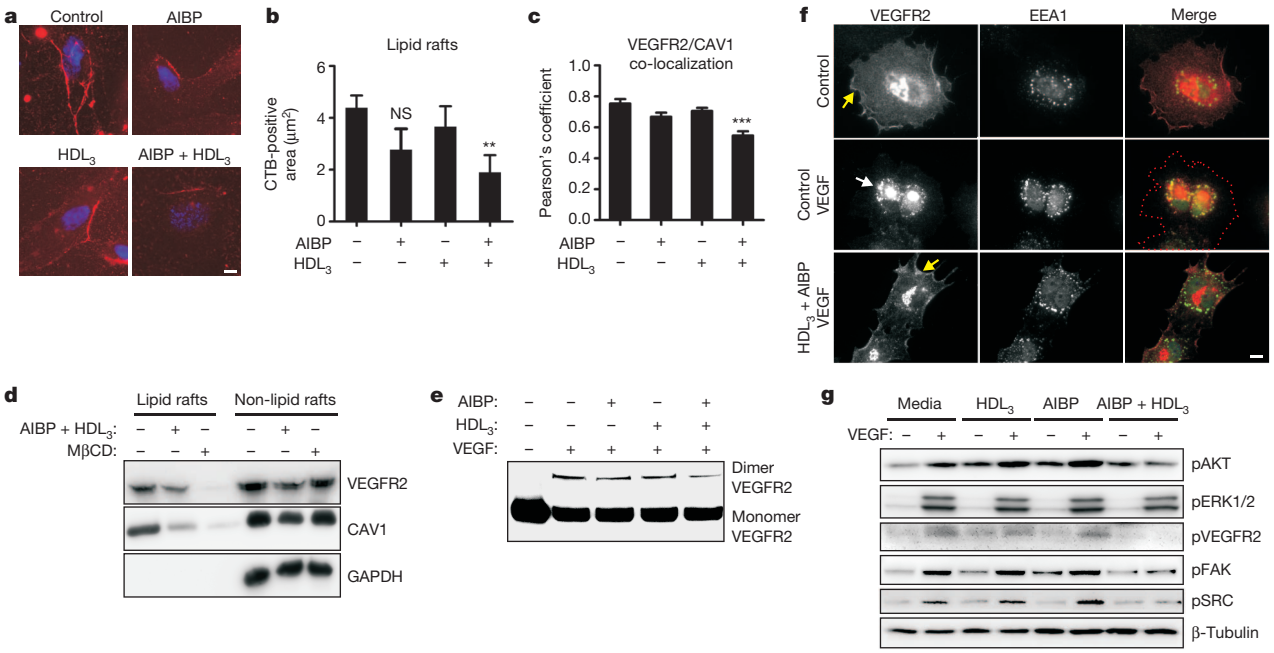


Figure 2 | Effect of AIBP on HUVEC lipid rafts, VEGFR2 localization, dimerization and signalling. **a**, Effect of human AIBP and HDL_3 on lipid rafts. HUVECs were pre-incubated with 50 μg ml^{-1} HDL_3, 0.1 μg ml^{-1} AIBP or 50 μg ml^{-1} HDL_3 + 0.1 μg ml^{-1} AIBP for 4 h. Cells were stained for nuclei (blue, 4′,6-diamidino-2-phenylindole (DAPI)) and for lipid rafts (red, cholera toxin B (CTB) + anti-CTB antibody). Scale bar, 10 μm. **b**, The area of lipid rafts per cell. Mean ± s.e.; $n = 10$; $^{**}P < 0.01$; NS, $P = 0.08$. **c**, Effect of human AIBP and HDL_3 on CAV1 and VEGFR2 surface localization. HUVECs were incubated with AIBP and/or HDL_3 as in **a**, fixed and stained with antibodies to CAV1 and VEGFR2. Images were captured using TIRF microscopy (Supplementary Fig. 6) and Pearson's coefficient was calculated to assess surface co-localization of VEGFR2 with CAV1. Mean ± s.e.; $n = 38$–50; $^{***}P < 0.001$. **d**, VEGFR2 and CAV1 localization to lipid rafts. HUVECs were incubated with 20 μg ml^{-1} cholesterol–MβCD for 6 h, followed by a 1-h incubation with or without 50 μg ml^{-1} HDL_3 + 0.1 μg ml^{-1} AIBP, or a 30-min incubation with 10 mM MβCD. HUVEC lysates were separated into lipid rafts and non-lipid rafts fractions by ultracentrifugation, run on SDS–PAGE and blotted with VEGFR2 and CAV1 antibodies. **e**, Effect of human AIBP and HDL_3 on VEGFR2 dimerization. HUVECs were pre-incubated with HDL_3 and/or AIBP as in **a**, followed by a 20-min stimulation with 50 ng ml^{-1} VEGF. Cells were treated with a crosslinking reagent, lysed and immunoprecipitated with a VEGFR2 antibody. Monomers and crosslinked dimers of VEGFR2 were visualized on western blot. **f**, Effect of human AIBP and HDL_3 on VEGFR2 endocytosis. HUVECs were pre-incubated with or without 50 μg ml^{-1} HDL_3 + 0.1 μg ml^{-1} AIBP for 4 h, then stimulated with 50 ng ml^{-1} VEGF for 20 min, fixed and stained with antibodies to VEGFR2 (red) and the early endosome marker EEA1 (green). Yellow and white arrows point to the surface and endosomal localization of VEGFR2. Red dotted line traces cell contour. Scale bar, 10 μm. **g**, Effect of human AIBP and HDL_3 on VEGFR2 signalling. HUVECs were pre-incubated with HDL_3 and/or AIBP as in **a**, followed by a 20-min stimulation with 50 ng ml^{-1} VEGF. Total cell lysates were run on SDS–PAGE and probed as indicated.

reversed inhibition of VEGFR2, FAK and AKT phosphorylation in AIBP/HDL_3-treated cells (Supplementary Fig. 11). Consistent with the effect on VEGF signalling, HUVEC migration towards a VEGF cue was significantly reduced in AIBP/HDL_3-treated cells (Supplementary Fig. 12). These results indicate that AIBP facilitates cholesterol efflux from HUVECs to HDL and that cholesterol depletion of the plasma membrane disrupts lipid rafts and VEGF signalling and inhibits VEGF-induced angiogenesis.

AIBP is evolutionary conserved between *Drosophila*, zebrafish, mouse and human (Supplementary Fig. 13a). Zebrafish have two genes, here termed *aibp1* (also known as *apoa1bp*) and *aibp2* (also known as *yjefn3*), that code for the Aibp1 and Aibp2 proteins, respectively (Supplementary Fig. 13b). The *aibp2* expression in 24–36 h post-fertilization (h.p.f.) zebrafish embryos shows a clear segmental pattern, co-localizing with the somite marker *myod1* (Fig. 3a and Supplementary Fig. 14). By 48 h.p.f., when segmental angiogenesis is completed, *aibp2* is no longer expressed in somites.

Both zebrafish Aibp2 and zebrafish Aibp1 bound to human apoA-I and to the HDL in human plasma, but only Aibp2 was effective in promoting cholesterol efflux from HUVECs to HDL_3 (Supplementary Figs 15, 16). Zebrafish embryos injected with antisense morpholino oligonucleotides targeting *aibp2* translation sites had increased levels of free (unesterified) cholesterol, whereas injections of *aibp1* or scrambled control morpholino oligonucleotide did not result in any changes (Fig. 3b, c and Supplementary Fig. 17). Thus, we focused on *aibp2*. Using the polarity-sensitive fluorescent probe Laurdan, we observed a higher membrane lipid order in the areas of growing segmental arteries (SeA) corresponding to tip cells compared to stalk cells (Fig. 3d, e), suggesting a higher content of lipid rafts in tip cells, which may positively regulate zebrafish Vegfr2 (encoded by gene *kdrl*) signalling. The membrane lipid order was increased in the SeA of *aibp2* morphants compared to controls, and the difference between tip and stalk cells was lost. To test the hypothesis that Aibp2-mediated cholesterol efflux regulates membrane order in growing SeA, we injected *aibp2* morphants with human HDL_3 or with bovine serum albumin (BSA). Adding an excess of HDL_3—to promote cholesterol efflux and to override the Aibp2 deficiency—annulled the increase in membrane order in SeA of *aibp2* morphants, and a spatially indiscriminate HDL_3 excess equalized the membrane order in tip and stalk cells. Adding an excess of BSA had no effect on the membrane order in *aibp2* morphants. Lysates of *aibp2* knockdown embryos displayed increased phosphorylation of Vegfr2, Akt and Src, and decreased phosphorylation of Erk1 (Fig. 3f and Supplementary Figs 18, 19). These results suggest that zebrafish Aibp2 regulates cholesterol levels, the membrane lipid order and Vegfr2 signalling and, thus, may control angiogenesis.

Indeed, injection of morpholino oligonucleotides targeting *aibp2* translation or splicing sites into one-cell stage embryos of *Tg(fli1:egfp)*^{y1} zebrafish, which express eGFP in endothelial cells[19], resulted in remarkable dysregulation of angiogenesis, with profound ectopic branching of SeA and sprouting of subintestinal veins (SIV) (Fig. 3g and Supplementary Figs 20, 21). The *aibp2* knockdown was validated in western blot (Fig. 3f). The ectopic branching of SeAs in *aibp2* morphants was partially rescued by forced expression of *aibp2* mRNA lacking the morpholino oligonucleotide target site (Fig. 3h).

The *aibp2* expression pattern (Fig. 3a) resembles that of type 3 semaphorins, non-cell-autonomous repellent cues that guide the patterning of developing SeAs through endothelial-specific Plexin D1 receptors[20]. To determine the cell autonomy of the Aibp2 effect on angiogenesis, we performed cell-transplantation experiments, using *Tg(fli1:egfp)*^{y1} donors. Fluorescent endothelial cells from wild-type donors found in non-fluorescent *aibp2* morphants displayed excessive branching and filopodial projections, whereas fluorescent endothelial cells from *aibp2* morphant donors found in wild-type recipients had normal morphology (Fig. 4a, b). In a gain-of-function experiment, overexpression of Aibp2 inhibited SeA sprouting from the dorsal aorta

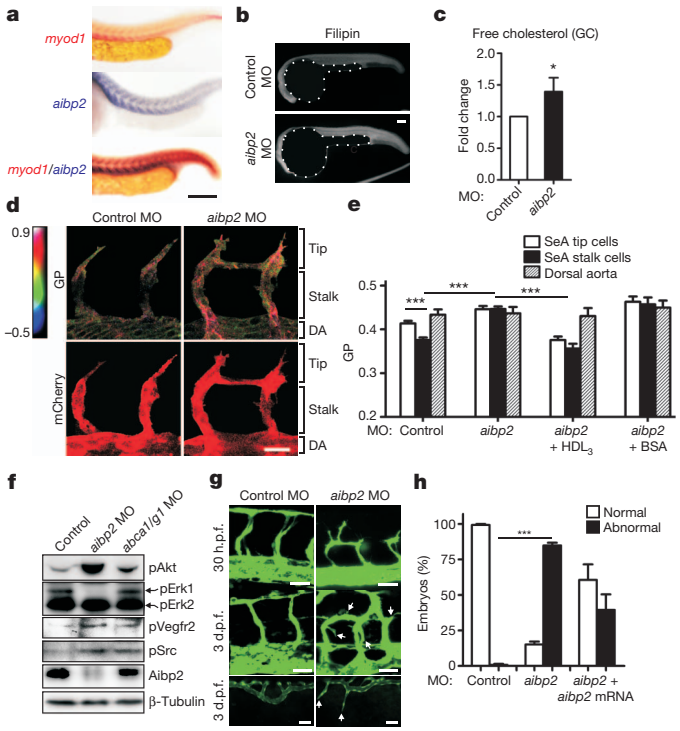


Figure 3 | Effect of Aibp deficiency on zebrafish cholesterol, membrane lipid order, Vegfr2 signalling and angiogenesis. **a**, Tissue distribution of *aibp2* mRNA in zebrafish embryos. Embryos at 24 h.p.f. were fixed and whole-mount *in situ* hybridization (WISH) was performed with antisense *myod1* and *aibp2* probes. Scale bar, 100 μm. **b**, **c**, Free cholesterol levels in *aibp2* morphants. **b**, Zebrafish embryos were injected with 8 ng of either control morpholino oligonucleotides (MO) or *aibp2* morpholino oligonucleotides. 24 h.p.f. control and *aibp2* morphants were stained with filipin to detect free cholesterol in embryos. Note the yolks are artificially masked on the images. **c**, At 24 h.p.f., the trunk area (without yolk) was dissected, total lipids extracted and free cholesterol levels determined by gas chromatography (GC). The cholesterol levels were normalized to the protein content and then to the values in control morpholino oligonucleotide embryos. Fifty to seventy embryos were pooled for each sample. Mean ± s.e.; $n = 4$; $^*P < 0.05$. **d**, Effect of *aibp2* morpholino oligonucleotides on SeA membrane lipid order. *Tg(flk1:ras-cherry)*^{s896} embryos were injected with control or *aibp2* morpholino oligonucleotides as in **b** and at 24 h.p.f. were stained with 5 μM Laurdan. In the same embryos, confocal images of mCherry fluorescence (bottom images) and the multiphoton images of Laurdan fluorescence (ex., 800 nm; em., 400–460 nm; 470–530 nm) were captured. The multiphoton results (top row images) are displayed as pseudo-coloured generalized polarization (GP, a measure of the membrane lipid order) images, cropped to show only the vasculature, that is, mCherry-positive areas. Scale bar, 20 μm. **e**, The graph shows GP values in the areas corresponding to tip and stalk cells of growing SeA and the dorsal aorta (DA) as indicated in **d**. Some one-cell-stage embryos were co-injected with 1 nl of 10 mg ml^{-1} human HDL_3 or BSA. Note the *y*-axis scale is from 0.2 to 0.5. Mean ± s.e.; $n = 44$–119 SeA in 25–49 embryos; $^{***}P < 0.001$. **f**, Phosphorylation of signalling proteins. Lysates of 24 h.p.f. control (8 ng control morpholino oligonucleotide), *aibp2* (8 ng *aibp2* morpholino oligonucleotide) and *abca1 abcg1* (4 ng *abca1* morpholino oligonucleotide + 4 ng *abcg1* morpholino oligonucleotide) morphants were separated on SDS–PAGE and immunoblotted as indicated. **g**, Angiogenic defects in *aibp2* morphants. One-cell-stage *Tg(fli1:egfp)*^{y1} zebrafish embryos were injected with 8 ng of either control or *aibp2* morpholino oligonucleotide. The images are of SeA in 30 h.p.f. embryos (top row), and of SeA (middle row) and of SIV (bottom row) in 3 days post-fertilization (d.p.f.) embryos. Arrows point to dysregulated sprouts. Scale bar, 25 μm. **h**, Quantification of the number of embryos with normal and abnormal angiogenesis (SeA with ectopic branching). The abnormal angiogenesis was partially rescued by co-injection of 40 pg of *aibp2* mRNA lacking the morpholino oligonucleotide-targeting site. Mean ± s.e.; $n = 100$–149. $^{***}P < 0.001$.

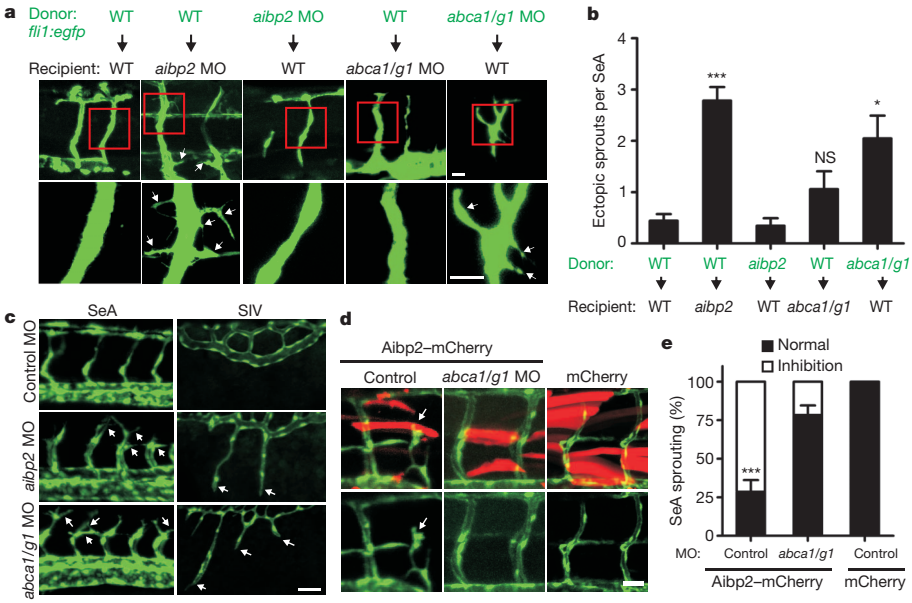


Figure 4 | Effect of Aibp and Abca1 Abcg1 deficiency on zebrafish angiogenesis. **a**, Mosaic expression analysis of endothelial cell branching in control, *aibp2* and *abca1 abcg1* knockdown embryos. At 4 h.p.f., cells were isolated from donor embryos and transplanted into recipient embryos. Recipient embryos were analysed at 3 d.p.f. Arrows point to aberrant ectopic branches/sprouts. Scale bar, 25 μm. **b**, Numbers of ectopic branches/filopodial projections per SeA. Mean ± s.e.; $n = 8$–16. $^{*}P < 0.05$, $^{***}P < 0.001$. **c**, Angiogenic defects in *abca1 abcg1* morphants. One-cell-stage embryos were injected with 8 ng of control morpholino oligonucleotide, 8 ng *aibp2* morpholino oligonucleotide or 4 ng *abca1* morpholino oligonucleotide + 4 ng *abcg1* morpholino oligonucleotide. Images of SeA (30 h.p.f.) and SIV (3 d.p.f.) are shown. Scale bar, 50 μm. **d**, Knockdown of *abca1* and *abcg1* cancels the effect of zebrafish Aibp2 overexpression. One-cell-stage embryos were injected with 2 nl of 100 ng μl^{-1} *myog:aibp2–mCherry*, *myog:aibp2–mCherry* plus *abca1* and *abcg1* morpholino oligonucleotides, or *myog:mCherry*. The arrow points to an aberrant SeA at the site of Aibp2–mCherry expression. Scale bar, 20 μm. **e**, Abnormal SeA formation was quantified in 8–16 embryos per group. Mean ± s.e.; $^{***}P < 0.001$.

and normal growth of sprouted SeA (Supplementary Fig. 22). These results suggest a role for Aibp2 as a repellent molecule whose function depends on the milieu surrounding endothelial cells but not on Aibp2 expression in the endothelial cells themselves.

The loss of zebrafish Aibp2 resulted in increased expression of genes involved in angiogenesis, such as *tie2*, *vegfr2*, *vegfr3* and *fli1* (Supplementary Figs 23, 24). Thus, in addition to the effect of Aibp2 on the membrane lipid order and Vegfr2 signalling (Figs 3d–f), Aibp2 also affects expression of angiogenic genes.

To further validate that effective cholesterol efflux is required for normal angiogenesis, we knocked down zebrafish cholesterol transporters *abca1* and *abcg1* (refs 21, 22) and observed higher levels of free cholesterol, increased levels of phosphorylated Akt, Vegfr2 and Src, and dysregulated SeA and SIV angiogenesis (Figs 3f, 4c and Supplementary Figs 19, 25–27), closely reproducing the angiogenesis defects of *aibp2* morphants. Individual knockdown of both *abca1* and *abcg1* suggested a dominant role of *abca1* in embryonic angiogenesis (Supplementary Fig. 28). In contrast to the *aibp2* non-cell-autonomous regulation of angiogenesis, fluorescent endothelial cells from *abca1 abcg1* morphant donors found in wild-type recipients displayed excessive SeA branching (Figs 4a, b), confirming that cholesterol efflux from endothelial cells is required to restrain ectopic angiogenesis. Overexpression of Aibp2–mCherry in somites resulted in inhibition of SeA growth, which was rescued by knocking down *abca1* and *abcg1* (Figs 4d, e). These results provide additional evidence that expression of zebrafish Aibp2 limits blood vessel growth and also suggest that zebrafish Abca1- and/or Abcg1-mediated cholesterol efflux is required for the effect of Aibp2 on angiogenesis.

On the basis of our results, we propose that there is an additional level of paracrine regulation of the VEGFR2 pathway in which cholesterol efflux and associated reduction of ordered membrane microdomains/lipid rafts interfere with the VEGFR2 membrane localization, dimerization, endocytosis and signalling. Because in 24 h.p.f. zebrafish *aibp2* mRNA is highly expressed in somites, but not in the inter-somitic spaces where SeA grow, it is likely that zebrafish Aibp2-mediated cholesterol efflux inhibits Vegfr2 signalling in a site-specific manner to prevent lateral protrusions from stalk and tip cells and restrains ectopic SeA growth into somites (Supplementary Fig. 1).

The role of cholesterol-efflux mechanisms in protecting against endothelial dysfunction, in particular in hypercholesterolemic animals prone to the development of atherosclerosis, has been reported[10,23]. However, our study is the first to demonstrate the role of AIBP in promoting cholesterol efflux from endothelial cells to HDL and the importance of this mechanism in regulation of angiogenesis. In contrast to the apoA-I-containing HDL, apoB-containing low-density lipoprotein and very-low-density lipoprotein deliver cholesterol and other lipids to the cell and, thus, are positioned to promote angiogenesis. Interestingly, a recent paper finds the opposite; that apoB lipoproteins negatively regulate angiogenesis in zebrafish embryos[24]. The authors suggest a mechanism in which the apoB protein, but not the lipid components within apoB-containing lipoproteins, is responsible for transcriptional regulation of Vegfr1, a soluble decoy receptor for VEGF. Our experiments uncovered a different, lipid-mediated mechanism in which effective cholesterol efflux is a critical process that ensures proper angiogenesis and Aibp secreted by the surrounding tissues serves as an important negative regulator of angiogenesis.

METHODS SUMMARY

Human and zebrafish AIBP were used in cholesterol efflux, angiogenesis and cellular assays as described[10,23,25]. *In vivo* angiogenesis was monitored in *Tg(fli1:egfp)*^{y1} zebrafish with *aibp2* and/or *abca1* and *abcg1* knockdown or forced *aibp2* expression.

Full Methods and any associated references are available in the online version of the paper.

Received 14 February 2012; accepted 8 April 2013.
Published online 29 May 2013.

1. Bodzioch, M. *et al.* The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nature Genet.* **22,** 347–351 (1999).
2. Rust, S. *et al.* Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1. *Nature Genet.* **22,** 352–355 (1999).

3. Klucken, J. *et al.* ABCG1 (ABC8), the human homolog of the *Drosophila white* gene, is a regulator of macrophage cholesterol and phospholipid transport. *Proc. Natl Acad. Sci. USA* **97,** 817–822 (2000).
4. Yvan-Charvet, L. *et al.* ATP-binding cassette transporters and HDL suppress hematopoietic stem cell proliferation. *Science* **328,** 1689–1693 (2010).
5. Armstrong, A. J., Gebre, A. K., Parks, J. S. & Hedrick, C. C. ATP-binding cassette transporter G1 negatively regulates thymocyte and peripheral lymphocyte proliferation. *J. Immunol.* **184,** 173–183 (2010).
6. Bensinger, S. J. *et al.* LXR signaling couples sterol metabolism to proliferation in the acquired immune response. *Cell* **134,** 97–111 (2008).
7. Ritter, M. *et al.* Cloning and characterization of a novel apolipoprotein A-I binding protein, AI-BP, secreted by cells of the kidney proximal tubules in response to HDL or apoA-I. *Genomics* **79,** 693–702 (2002).
8. Jha, K. N. *et al.* Biochemical and structural characterization of apolipoprotein A-I binding protein, a novel phosphoprotein with a potential role in sperm capacitation. *Endocrinology* **149,** 2108–2120 (2008).
9. Stefulj, J. *et al.* Human endothelial cells of the placental barrier efficiently deliver cholesterol to the fetal circulation via ABCA1 and ABCG1. *Circ. Res.* **104,** 600–608 (2009).
10. Terasaka, N. *et al.* ABCG1 and HDL protect against endothelial dysfunction in mice fed a high-cholesterol diet. *J. Clin. Invest.* **118,** 3701–3713 (2008).
11. Fessler, M. B. & Parks, J. S. Intracellular lipid flux and membrane microdomains as organizing principles in inflammatory cell signaling. *J. Immunol.* **187,** 1529–1535 (2011).
12. Mendez, A. J. *et al.* Membrane lipid domains distinct from cholesterol/sphingomyelin-rich rafts are involved in the ABCA1-mediated lipid secretory pathway. *J. Biol. Chem.* **276,** 3158–3166 (2001).
13. Murphy, A. J. *et al.* High-density lipoprotein reduces the human monocyte inflammatory response. *Arterioscler. Thromb. Vasc. Biol.* **28,** 2071–2077 (2008).
14. Noghero, A. *et al.* Liver X receptor activation reduces angiogenesis by impairing lipid raft localization and signaling of vascular endothelial growth factor receptor-2. *Arterioscler. Thromb. Vasc. Biol.* **32,** 2280–2288 (2012).
15. Oshikawa, J. *et al.* Novel role of p66Shc in ROS-dependent VEGF signaling and angiogenesis in endothelial cells. *Am. J. Physiol. Heart Circ. Physiol.* **302,** H724–H732 (2012).
16. Ikeda, S. *et al.* Novel role of ARF6 in vascular endothelial growth factor-induced signaling and angiogenesis. *Circ. Res.* **96,** 467–475 (2005).
17. Liao, W. x. *et al.* Compartmentalizing VEGF-induced ERK2/1 signaling in placental artery endothelial cell caveolae: a paradoxical role of caveolin-1 in placental angiogenesis *in vitro*. *Mol. Endocrinol.* **23,** 1428–1444 (2009).
18. Eichmann, A. & Simons, M. VEGF signaling inside vascular endothelial cells and beyond. *Curr. Opin. Cell Biol.* **24,** 188–193 (2012).
19. Lawson, N. D. & Weinstein, B. M. *In vivo* imaging of embryonic vascular development using transgenic zebrafish. *Dev. Biol.* **248,** 307–318 (2002).
20. Torres-Vázquez, J. *et al.* Semaphorin-plexin signaling guides patterning of the developing vasculature. *Dev. Cell* **7,** 117–123 (2004).
21. Dean, M. & Annilo, T. Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annu. Rev. Genomics Hum. Genet.* **6,** 123–142 (2005).
22. Archer, A. *et al.* Transcriptional activity and developmental expression of liver X receptor (*lxr*) in Zebrafish. *Dev. Dyn.* **237,** 1090–1098 (2008).
23. Whetzel, A. M. *et al.* ABCG1 deficiency in mice promotes endothelial activation and monocyte–endothelial interactions. *Arterioscler. Thromb. Vasc. Biol.* **30,** 809–817 (2010).
24. Avraham-Davidi, I. *et al.* ApoB-containing lipoproteins regulate angiogenesis by modulating expression of VEGF receptor 1. *Nature Med.* **18,** 967–973 (2012).
25. Carmona, G. *et al.* Role of the small GTPase Rap1 for integrin activity regulation in endothelial cells and angiogenesis. *Blood* **113,** 488–497 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Traver, N. Chi, J. Witztum, R. Klemke, D. Yelon, T. Handel, K. Stoletov, W. Clements, C. Pouget, Z. Garavito-Aguilar, A. Ablooglu, R. Zhang, X. Yang, M. Angert, K. Pestonjamasp and J. Santini (University of California, San Diego), C. Hedrick, K. Ley, D. Sag, P. Sundd and A. Wu (La Jolla Institute for Allergy and Immunology), S. Trzaska (New York University), S. J. Du (University of Maryland), B. Schmid and C. Haass (Ludwig-Maximilians-University München), D. Owen and A. Magenau (University of New South Wales), A. Siekmann (Max Planck Institute for Molecular Biomedicine) and C. Binder (Medical University of Vienna) for discussions, technical assistance and/or for providing reagents and access to equipment for this study. The project was supported by the NIH grants HL093767 (Y.I.M.), HL055798 (Y.I.M.) and HL114734 (L.F.), and the fellowship 18FT-0137 from the UC Tobacco-Related Disease Program (L.F.), as well as the UCSD Neuroscience Microscopy Facility Grant P30 NS047101. The authors declare no conflicts of interests.

Author Contributions L.F. and Y.I.M. conceived the project, designed the experiments and wrote the manuscript. J.T.-V. made important intellectual contributions and helped revise the manuscript. L.F. performed the majority of the experiments. S.-H.C., J.S.B., C.L., F.A., F.U., P.W., A.T., E.D., J.P., A.C.L. performed experiments and/or provided research assistance.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.I.M. (yumiller@ucsd.edu).

METHODS

Cloning of human and zebrafish AIBP, recombinant protein expression and purification and antibody production. Zebrafish *aibp2* (NCBI Gene ID:557840) and *aibp1* (NCBI Gene ID:436891) were cloned from zebrafish brain complementary DNA using primers: CCGGAATTCCATGTTGGGGGTTCGAGCTCTG (5′) and CGCGGATCCTCAGTTGAGCTGAAACACACACTC (3′) for *aibp1* and CCGGAATTCCGCCACCATGAACCACAGCTCCAACG (5′) and CGCGGATCCCGCAGTTCTATAATACATTCTGTGC (3′) for *aibp2*. The fragments were cloned in frame into pFLAG-CMV4 (Sigma). Human *APOA1BP* (Gene ID: 128240) was cloned from HEK293 cell cDNA using primers: CCGGAATTCCATGTCCAGGCTGCGGGCGCTGCTGGGCCTCG (5′) and CGGGGTACCTCACTGCAGACGATAGACACACTC (3′). For expression of AIBP proteins, the genes were cloned in frame into pHUE vector[26] (provided by T. Handel), expressed in BL21 DE3 competent cells (Invitrogen) and purified with a Ni-NTA agarose resin column (Qiagen). Deubiquitinase (DUB) expressed in pHUE was used as a negative control in experiments with recombinant AIBP. To produce a zebrafish Aibp2 antibody, recombinant Aibp2 was mixed with complete Freund's adjuvant (Sigma) and injected subcutaneously into a guinea pig. The guinea pig was boosted three more times. Post-immune plasma was compared with pre-immune plasma from the same animal and used in western blot to detect Aibp2 in zebrafish lysates (Supplementary Fig. 18). The specificity of the antibody was confirmed by adding excess of recombinant Aibp2 to the antibody, which prevented its binding to a specific band on the western blot.

Cholesterol efflux. A cholesterol efflux assay was performed as described[23,27], with modifications. In brief, HUVECs (ATCC) were loaded with 2 μCi ml^{-1} ^{3}H-cholesterol, and cholesterol efflux was initiated by the addition of 0.2% BSA/endothelial cell basal medium (EBM) with 50 μg ml^{-1} HDL_3 (isolated from normolipidemic human plasma by ultracentrifugation), in the presence or absence of 0.2 μg ml^{-1} zebrafish Aibp2, Aibp1, or human AIBP. DUB, replacing AIBP, was used as a negative control. Background, nonspecific release of ^{3}H-cholesterol was measured in absence of HDL. After 1–6 h of incubation, the medium was collected and counted in a liquid scintillation counter LS 6500 (Beckman Coulter). The cells were extracted with 2-propanol, and the lipid extract was added to ScintiVerse BD Cocktail (Fisher) and counted. Cholesterol efflux was expressed as a percentage of ^{3}H counts in the medium compared to combined ^{3}H counts in the cells and the medium. Background, nonspecific release of ^{3}H from the cells was subtracted.

***ABCG1* knockdown.** Both negative control and *ABCG1* siRNA oligonucleotides were from Ambion. HUVECs were plated in 6-well plates at 5×10^5 cells per well and transfected with 66.6 nM siRNA using SuperFect Transfection Reagent (Qiagen) as described in the manufacturer's protocol. Two days after transfection, cells were washed and used in an efflux assay. Two additional wells of transfected cells were used to confirm *ABCG1* knockdown in western blot using an antibody from Novus Biologicals.

AIBP/HDL_3–HUVEC binding assay. Human AIBP and HDL_3 were biotinylated with EZ-Link Sulfo-NHS-Biotin (Thermo Scientific) according the manufacturer's protocol. Binding of biotinylated human AIBP or biotinylated HDL_3 to HUVECs was assessed by a chemiluminescent binding assay as described in ref. 28, with modifications. HUVECs (2×10^4) were seeded into 96-well flat bottom plates in 5% FBS-EBM. After 72 h, plates were blocked with ice-cold 1% BSA-PBS for 30 min on ice, incubated with ice-cold biotinylated proteins for 2 h on ice, washed, and fixed with ice-cold 4% paraformaldehyde (PFA) in PBS for 30 min. HUVEC-bound biotinylated human AIBP or HDL_3 were detected with NeutrAvidin-conjugated alkaline phosphatase (Pierce) and LumiPhos 530 (Lumigen), using a Dynex luminometer (Dynex Technologies). Data were recorded as relative light units counted per 100 ms. All samples were assayed in triplicates. The parameters of human AIBP and HDL_3 binding to HUVECs (B_{max} and K_d) were calculated using a total and nonspecific binding algorithm within the GraphPad Prism 5.0 software package. The following model was used: H + C ↔ HC, where H is unbound HDL_3, C is cells, and HC is HDL_3 bound to the cells. The equations used for calculating binding parameters were:

$$[HC]_{specific} = B_{max} \times [H]/([H] + K_d)$$

$$[HC]_{nonspecific} = a + b \times [H]$$

$$[HC]_{total} = [HC]_{specific} + [HC]_{nonspecific},$$

where a is background and b is the slope of the linear fit of nonspecific binding. Goodness of fit of nonlinear regression was estimated using R^2 and standard deviation of residuals (Sy.x), expressed in the same units as [H] and B_{max}. A molecular mass of 80 kDa was used for the HDL protein.

***In vitro* angiogenesis assay.** The angiogenesis assay was carried out as described in refs 25 and 29. Growth factor reduced Matrigel (BD Biosciences) was thawed at 4 °C overnight and diluted with an equal volume of serum-free EBM medium (Lonza). Each well of 96-well plates was coated with 50 μl diluted Matrigel and incubated at 37 °C for 1 h. HUVECs were serum-starved and then pre-incubated with HLD_3 and/or human AIBP. Cells were collected and added to Matrigel-coated plates at 1×10^4 cells per well in EBM, in the presence or absence of 20 ng ml^{-1} VEGF (R&D Systems). Following a 12-h incubation, tubular structures were imaged with a phase contrast microscope.

Free cholesterol measurements in HUVECs. HUVEC cholesterol levels were measured in cellular lipid extracts using a colorimetric assay (BioVision) as previously described[14].

Aortic ring neovascularization assay. The method was adopted from ref. 30, with modifications. Thoracic aorta was isolated from a 6-week-old male C57BL/6 mouse or an age- and gender-matched *Abcg1*$^{-/-}$ mouse (provided by C. Hedrick), cleaned from surrounding fat and connective tissue and sliced into 1-mm-long rings. The aortic rings were placed in wells of a 48-well plate containing solidified Matrigel and then covered with additional Matrigel. Small wells were made in Matrigel approximately 0.5 mm from aortic rings and 50-μl aliquots of Matrigel containing 1×10^5 HEK293 cells transfected with mCherry (negative control), zebrafish Aibp2 or human AIBP were placed in these wells. After 10 min, each well was filled with EBM medium supplemented with 10 ng ml^{-1} VEGF and the plates were incubated at 37 °C for 6 days. Media were changed every 2 days. The rings were photographed in phase contrast using a Nikon Eclipse Ti microscope.

Visualization of lipid rafts with cholera toxin B. HUVECs were plated on glass coverslips and pre-incubated with 50 μg ml^{-1} HDL_3, 0.1 μg ml^{-1} human AIBP or 50 μg ml^{-1} HDL_3 plus 0.1 μg ml^{-1} human AIBP for 4 h. Cells were washed once with medium before the addition of 1 μg ml^{-1} Alexa Fluor 594-labelled CTB (Invitrogen). Cells were incubated for 15 min at 4 °C, washed with PBS and then incubated for 15 min at 4 °C with an anti-CTB antibody (EMD Chemicals) to crosslink CTB and lipid rafts. After washing with PBS, cells were fixed in 4% PFA for 20 min at 4 °C, mounted with a Prolong Antifade Kit with DAPI (Invitrogen) and images were captured with a Leica DM IRE2 fluorescent microscope.

Cell fractionation. Lipid rafts (light membrane fractions) were isolated using a detergent-free, discontinuous gradient ultracentrifugation method[14]. In brief, HUVECs were washed twice with ice-cold PBS and cells were scraped from the plate in 0.5 M sodium carbonate buffer (pH 11.0) containing a protease inhibitor cocktail (Sigma), homogenized and sonicated three times for 10 s. Samples were adjusted to 45% sucrose by adding a 90% sucrose solution in MBS (25 mM 2-(*N*-morpholino)ethanesulfonic acid, 0.15 M NaCl, pH 6.5) and placed into ultracentrifugation tubes. A 5–35% sucrose discontinuous gradient was formed above the sample, followed by ultracentrifugation at 35×10^3 r.p.m. for 18 h at 4 °C in a SW-41 rotor (Beckman). Ten 1-ml fractions were collected from the top to the bottom of each gradient. The lipid rafts fraction (fraction 5) and the non-lipid rafts fraction (fraction 10) were used for further analysis, which included measurements of protein concentration and immunoblotting. Thirty microlitres of lipid rafts and non-lipid rafts fractions (adjusted to load equal protein concentrations of each sample) were run on SDS–PAGE, transferred to polyvinylidene difluoride (PVDF) membranes and blotted with the indicated antibodies.

CAV1 and VEGFR2 co-localization. HUVECs plated on chamber coverglass (Lab-Tek II) were incubated for 4 h with 0.1 μg ml^{-1} human AIBP, 50 μg ml^{-1} HDL_3 or 0.1 μg ml^{-1} AIBP + 50 μg ml^{-1} HDL_3, and the cells were washed with PBS and fixed with warm 4% PFA for 15 min at room temperature (24 °C). HUVECs were permeabilized, blocked, and incubated with anti-CAV1 (BD Biosciences) and anti-VEGFR2 (Cell Signaling Technology) antibodies, followed by incubation with anti-mouse IgG-Alexa Fluor 488 and anti-rabbit IgG-Cy3 antibodies. Images were captured using a Nikon Eclipse Ti inverted fluorescent microscope operating in TIRF mode. Raw TIFF images of CAV1 and VEGFR2 were analysed using JACoP algorithm[31] in Image J. Co-localization was quantified using Pearson's coefficient.

VEGFR2 endocytosis. HUVECs were incubated for 4 h with 0.1 μg ml^{-1} human AIBP, 50 μg ml^{-1} HDL_3 or 0.1 μg ml^{-1} human AIBP + 50 μg ml^{-1} HDL_3, followed by a 20-min incubation with 50 ng ml^{-1} VEGF. Cells were fixed and stained with antibodies against VEGFR2 and the early endosomal marker EEA1-FITC (BD Biosciences). Images were captured with a Nikon Eclipse Ti inverted fluorescent microscope. VEGFR2 and EEA1 co-localization was quantified using Pearson's coefficient with the JACoP plugin loaded to ImageJ[31].

VEGFR2 dimerization assay. The assay was carried out as described in ref. 32. Two days after plating 1×10^6 HUVECs in a 10-cm dish, the cells were starved overnight in 0.5% FBS-EBM. The next day, cells were incubated with human AIBP and/or HDL_3, followed by a 20-min incubation with 50 ng ml^{-1} VEGF and then crosslinked with 1 mg ml^{-1} bis-sulfosuccinimidyl (Thermo Scientific) for 30 min on ice. Cell lysates were immunoprecipitated with an anti-VEGFR2 antibody

immobilized on agarose beads. The beads were washed and the eluted samples were run on SDS–PAGE, followed by immunoblotting with the VEGFR2 antibody.

HUVEC migration assay. Serum-starved HUVECs were pretreated with 50 μg ml^{-1} HDL_3, 0.2 μg ml^{-1} human AIBP or 50 μg ml^{-1} HDL_3 + 0.2 μg ml^{-1} human AIBP for 4 h at 37 °C in 5% lipoprotein-deficient serum (LPDS) and EBM (Lonza), collected from the plate, washed, re-suspended in 5% LPDS/EBM and added to the transwell (8 μm pore size). VEGF was added to the lower chamber at 20 ng ml^{-1}. Following a 4-h incubation, the transwell membranes were fixed in ice-cold methanol for 10 min and stained with filtered 0.5% Crystal Violet for 10 min, and transmigrated cells were counted.

Zebrafish. Wild-type AB and transgenic *Tg(fli1:egfp)^{y1}* and *Tg(flk1:ras-cherry)^{s896}* zebrafish lines[19,33] were provided by D. Traver and N. Chi (UCSD). Zebrafish were maintained as previously described[34], and all experimental procedures were approved by the UCSD Institutional Animal Care and Use Committee.

Confocal microscopy. Confocal imaging was carried out as previously described[35]. In brief, anaesthetized zebrafish embryos (treated at 24 h.p.f. with 0.003% 1-phenyl 2-thiourea) were housed in a sealed chamber (Invitrogen) in a small drop of 0.02% tricaine (Sigma) containing E3 embryonic medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$ and 0.1% methylene blue) and imaged using a Nikon C1si confocal microscope. *Z*-stacks were acquired with a 1–3-μm step, and images were three-dimensionally rendered and analysed using Imaris software (Bitplane). All three-dimensional reconstructions were performed with the same threshold settings.

Morpholino oligonucleotide injections. To knock down gene expression, 4–8 ng of antisense morpholino oligonucleotides (synthesized by GeneTools) were injected into one-cell-stage embryos. A control morpholino oligonucleotide was derived from the *aibp2* sequence, with five mismatched oligonucleotides. Control morpholino oligonucleotides, 5′-TGAGCTTCATGTTCATTTATTCCGC-3′; *aibp2* morpholino oligonucleotides (*aibp2* MO1), 5′-TGTGGTTCATCTTGATTATTCGGC-3′; *aibp2* splicing morpholino oligonucleotides (*aibp2* MO2), 5′-TGTTGAGTGTCAGACAAACCTTGGT-3′; *aibp1* morpholino oligonucleotides, 5′-TCTGTATTCAAATCAGACGCTCAGT-3′; *abca1* morpholino oligonucleotides, 5′-AACCCAACTGAGTGGAGACAGCCAT-3′; *abcg1* morpholino oligonucleotides, 5′-AAAAGGCTGCCATGAGACATGCCAT-3′.

Quantitative analysis of cell sprouts in SeAs and SIVs. To determine changes in segmental artery cell sprouts in embryos injected with morpholino oligonucleotides targeting *aibp2*, *abca1* and/or *abcg1*, we counted abnormal projections in 4 to 6 pairs of segmental arteries in adjacent somite boundaries in each zebrafish. For each set of injections, 15 embryos (that is, 60–90 sprouts) were examined. Values were expressed as a number of ectopic sprouts per SeA. To examine sprouts in SIV, only the sprouts moving in the ventral direction out of the SIV were counted. Values were expressed as a number of ventral SIV sprouts per zebrafish.

Measuring membrane lipid order with polarity-sensitive probe. The experiments were carried out as described in refs 36 and 37. In brief, live *Tg(flk1:ras-cherry)^{s896}* zebrafish embryos were incubated with 5 μM Laurdan (Invitrogen) at 28 °C for 30 min. The concentration of a Laurdan stock solution was measured using attenuance at 365 nm and an extinction coefficient of 19,000 $cm^{-1}M^{-1}$. After incubation with Laurdan, embryos were incubated with E3 medium for additional 30 min, fixed in PFA for 4 h at room temperature, de-yolked and embedded in 1% low-melting-temperature agarose for imaging. Images were captured with a Leica SP5 confocal/multiphoton system, using a water immersion ×20 objective. The confocal mode was used to capture mCherry fluorescence and the multiphoton mode was used to capture Laurdan images (ex., 800 nm, em., 400–460 nm and 470–530 nm) in the same embryos. The multiphoton results were displayed as pseudo-coloured GP (a measure of the membrane lipid order) images, derived from Laurdan ratiometric measurements and using a ImageJ plug-in as described[36]. The quantitative data were obtained by measuring GP values in the areas corresponding to tip cells (top one-third of the SeA length), stalk cells (bottom two-thirds of the SeA length) and the dorsal aorta in several individual, mCherry-masked *z*-sections. This method ensured that GP values were derived only from endothelial cells (for example, from the areas where mCherry was in focus in each *z*-section). The GP images in Fig. 3d were composed each from four to five individual *z*-sections, with a minimal overlap of mCherry-masked GP images.

WISH. WISH was carried out as described[38–40]. In brief, wild-type embryos or morphants at indicated developmental stages were fixed with 4% PFA and the embryos older than 24 h.p.f. were permeabilized with 10 μg ml^{-1}proteinase K (Roche). Subsequently, the embryos were pre-hybridized at 70 °C for 4–6 h, and then hybridized with a digoxigenin-labelled *aibp2* antisense probe at 65 °C for 2 days. Both the control sense and anti-sense RNA probes were directly synthesized from *aibp2* full-length gene using a Roche T7/SP6 RNA or Ambion T3 RNA synthesis kit. After extensive wash, hybridized RNA was detected by immunohistochemistry using an alkaline-phosphatase-conjugated antibody against digoxigenin (Roche) and a chromogenic substrate nitro blue tetrazolium (NBT) (Sigma) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP)[38] (Sigma). A similar procedure was performed with *tie2*, *vegfr2*, *vegfr3*, *fli1* and *cdh5* probes. Double WISH was performed as described[41]. Both digoxigenin-labelled *aibp2* and fluorescein-labelled *myod1* were hybridized with the same embryos. The embryos were then first incubated with alkaline-phosphatase-conjugated anti-fluorescein antibody, and fast red (Roche) was used as a substrate. Subsequently, the embryos were washed, fixed with 4% PFA and incubated with anti-digoxigenin antibody conjugated alkaline phosphatase, and then NBT/BCIP was used as a chromogenic substrate.

Transplantation experiments. Cell transplantation was performed as described[39], with different combinations of donors and recipients as indicated in Fig. 4a. Donor embryos were of the *Tg(fli1:egfp)^{y1}* origin, and recipient embryos were of the wild-type origin. At the one-cell stage, donor *Tg(fli1:egfp)^{y1}* embryos were injected with rhodamine-labelled dextran (mini-ruby, Invitrogen) as a lineage tracer. At the sphere stage (approximately 4 h.p.f.), embryos were dechorionated by 0.4 mg ml^{-1} pronase (Sigma) and transferred to agarose wells (Adaptive Science Tools). Approximately 20–40 cells from the margin of a donor embryo were transferred to the margin of a recipient embryo. The recipient embryos were subsequently grown at 28 °C and imaged at 72 h.p.f. Endothelial cells in chimaeric zebrafish originating from donor embryos were visualized by their green fluorescence using a Leica M165FC fluorescent stereoscope. For detailed analysis, images were captured using a Nikon C1si confocal microscope. Numbers of ectopic branches in each fluorescent SeA were counted.

Real-time PCR. Real-time PCR was performed using a Rotor Gene Q PCR machine (Qiagen). Real-time PCR master mix Platinum SYBR Green qPCR SuperMix was from Invitrogen. The primers were synthesized by Integrated DNA Technologies. The PCR program was: 50 °C for 2 min (UDG incubation), 95 °C for 2 min, 40 cycles of: 95 °C for 15 s, 60 °C for 1 min. Primer sequences: *fli1*: forward CTTGGCACGTTGCCTTGATAAG, reverse CCTTCATATCTGAGAGTGATCCC; *tie2*: forward GCGATGGATGGCAATAGAGT; reverse CGACAGCAGGATCTGAGAGA; *vegfr2*: forward TCCACGAGGGTGGGCAGTCA, reverse AGACGGGTGGTGTGGAGTAACGA; *kdrb*: forward TGCCCACATGGAGCTGCTAGCA, reverse TGTGGCACATTCAACCACATGAGC; *actb1*: forward CTCTTCCAGCCTTCCTTCCT, reverse GGTTGGTCGTTCGTTTGAAT.

Immunoblot of zebrafish lysates. Zebrafish were lysed on ice with a lysis buffer (50 mM Tris-HCl, pH 7.5, 4 mM sodium deoxycholate, 1% Triton X-100, 150 mM NaCl, 1 mM EDTA and a protease inhibitor cocktail from Sigma). Protein content was determined with a DC protein assay kit (BioRad) and equal protein amounts of the cell lysates were run on a 4–12% Bis-Tris SDS–PAGE with MOPS buffer (Invitrogen) and then transferred to a PVDF membrane (Invitrogen). The blots were probed with appropriate antibodies against specific phosphorylated and non-phosphorylated proteins (Cell Signaling Technology), secondary antibodies conjugated with horseradish peroxidase and developed using a SuperSignal West Dura substrate (Pierce).

Filipin staining. Filipin staining of embryos was performed as described[42]. Zebrafish were fixed with 4% PFA overnight at 4 °C. The fixed fish were incubated overnight with 0.05% filipin (Sigma) in PBS with 1% sheep serum, and then washed three times with PBS. Images were captured immediately with a Leica M165FC fluorescent stereoscope and quantified.

Total lipid extraction and free cholesterol measurements in zebrafish. Total lipid was extracted from zebrafish embryos as we previously described[28]. In brief, trunk/tail segments were dissected from 50 24 h.p.f. embryos and pooled together. The tissue was homogenized and supplemented with 50 μg stigmasterol, an internal standard to control for recovery of extracted sterols. Total lipid extraction was performed with 1:2 methanol/dichloromethane. No saponification of cholesterol esters was performed because the goal of this study was to measure free cholesterol, the form of cholesterol transferred from the cells via ABC transporters to apoA-I/HDL. Cholesterol and stigmasterol were measured with a Shimadzu GC-2014 gas chromatograph using a 30 m × 0.25 mm (i.d.) ZB-5HT Inferno capillary column, film thickness 0.2 μm (Phenomenex). Cholesterol levels were normalized to protein and then to the levels in embryos injected with control morpholino oligonucleotides.

26. Catanzariti, A. M., Soboleva, T. A., Jans, D. A., Board, P. G. & Baker, R. T. An efficient system for high-level expression and easy purification of authentic recombinant proteins. *Protein Sci.* **13,** 1331–1339 (2004).
27. O'Connell, B. J. Cellular physiology of cholesterol efflux in vascular endothelial cells. *Circulation* **110,** 2881–2888 (2004).
28. Fang, L. *et al.* Oxidized cholesteryl esters and phospholipids in zebrafish larvae fed a high cholesterol diet: macrophage binding and activation. *J. Biol. Chem.* **285,** 32343–32351 (2010).
29. Gao, F. *et al.* L-5F, an apolipoprotein A-I mimetic, inhibits tumor angiogenesis by suppressing VEGF/basic FGF signaling pathways. *Integr. Biol.* **3,** 479–489 (2011).
30. Bellacen, K. & Lewis, E. C. Aortic ring assay. *J. Vis. Exp.* **33,** 1564 (2009).

31. Bolte, S. & Cordelieres, F. P. A guided tour into subcellular colocalization analysis in light microscopy. *J. Microsc.* **224,** 213–232 (2006).
32. Chung, T. W. *et al.* Ganglioside GM3 inhibits VEGF/VEGFR-2-mediated angiogenesis: direct interaction of GM3 with VEGFR-2. *Glycobiology* **19,** 229–239 (2009).
33. Chi, N. C. *et al.* Foxn4 directly regulates *tbx2b* expression and atrioventricular canal formation. *Genes Dev.* **22,** 734–739 (2008).
34. Westerfield, M. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio rerio)* 5th edn (Univ. of Oregon Press, 2007).
35. Stoletov, K. *et al.* Vascular lipid accumulation, lipoprotein oxidation, and macrophage lipid uptake in hypercholesterolemic zebrafish. *Circ. Res.* **104,** 952–960 (2009).
36. Owen, D. M., Rentero, C., Magenau, A., Abu-Siniyeh, A. & Gaus, K. Quantitative imaging of membrane lipid order in cells and organisms. *Nature Protocols* **7,** 24–35 (2012).
37. Gaus, K., Le Lay, S., Balasubramanian, N. & Schwartz, M. A. Integrin-mediated adhesion regulates membrane order. *J. Cell Biol.* **174,** 725–734 (2006).
38. Thisse, C. & Thisse, B. High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nature Protocols* **3,** 59–69 (2008).
39. Siekmann, A. F. & Lawson, N. D. Notch signalling limits angiogenic cell behaviour in developing zebrafish arteries. *Nature* **445,** 781–784 (2007).
40. Lawson, N. D. *et al.* Notch signaling is required for arterial-venous differentiation during embryonic vascular development. *Development* **128,** 3675–3683 (2001).
41. Jowett, T. Analysis of protein and gene expression. *Methods Cell Biol.* **59,** 63–85 (1998).
42. Schwend, T., Loucks, E. J., Snyder, D. & Ahlgren, S. C. Requirement of Npc1 and availability of cholesterol for early embryonic cell movements in zebrafish. *J. Lipid Res.* **52,** 1328–1344 (2011).

LETTER

doi:10.1038/nature12180

Structure-guided discovery of the metabolite carboxy-SAM that modulates tRNA function

Jungwook Kim[1], Hui Xiao[2], Jeffrey B. Bonanno[1], Chakrapani Kalyanaraman[3], Shoshana Brown[4], Xiangying Tang[1], Nawar F. Al-Obaidi[1], Yury Patskovsky[1], Patricia C. Babbitt[4], Matthew P. Jacobson[3], Young-Sam Lee[5] & Steven C. Almo[1,6]

The identification of novel metabolites and the characterization of their biological functions are major challenges in biology. X-ray crystallography can reveal unanticipated ligands that persist through purification and crystallization. These adventitious protein–ligand complexes provide insights into new activities, pathways and regulatory mechanisms. We describe a new metabolite, carboxy-*S*-adenosyl-L-methionine (Cx-SAM), its biosynthetic pathway and its role in transfer RNA modification. The structure of CmoA, a member of the SAM-dependent methyltransferase superfamily, revealed a ligand consistent with Cx-SAM in the catalytic site. Mechanistic analyses showed an unprecedented role for prephenate as the carboxyl donor and the involvement of a unique ylide intermediate as the carboxyl acceptor in the CmoA-mediated conversion of SAM to Cx-SAM. A second member of the SAM-dependent methyltransferase superfamily, CmoB, recognizes Cx-SAM and acts as a carboxymethyltransferase to convert 5-hydroxyuridine into 5-oxyacetyl uridine at the wobble position of multiple tRNAs in Gram-negative bacteria[1], resulting in expanded codon-recognition properties[2,3]. CmoA and CmoB represent the first documented synthase and transferase for Cx-SAM. These findings reveal new functional diversity in the SAM-dependent methyltransferase superfamily and expand the metabolic and biological contributions of SAM-based biochemistry. These discoveries highlight the value of structural genomics approaches in identifying ligands within the context of their physiologically relevant macromolecular binding partners, and in revealing their functions.

Transfer RNAs contain many post-transcriptional modifications; nearly 100 distinct modifications have been reported[1]. Nucleotides at the wobble position (that is, the 5′ nucleotide of the anticodon triplet) are the most frequent targets for such modifications, as they confer efficient and accurate pairing between anticodons and cognate codon sequences. For example, wobble uridines in Gram-negative bacteria are often modified at C5 to 5-oxyacetyl uridine (cmo5U) (Fig. 1a), which enables multiple tRNAs to decode four of their respective degenerate codons[3]. This expanded recognition results from structural and tautomeric constraints imposed by the 5-oxyacetyl modification[4].

Mutagenesis studies established that genes responsible for chorismate biosynthesis are necessary for cmo5U formation and it was shown that one carbon atom of the oxyacetyl group originates from SAM[5,6]. Gene-disruption studies established that two members of the SAM-dependent methyltransferase superfamily (SDMT), CmoA and CmoB, were required for cmo5U formation[2]. Inactivation of *cmoA* resulted in formation of incompletely modified tRNAs, and hydroxy uridine (ho5U) and methoxy uridine (mo5U) were observed instead of cmo5U. In *cmoB*-defective mutants, only ho5U was detected. Despite these observations, the roles of CmoA and CmoB in the transformation of ho5U to cmo5U remain unclear. Wobble uridine hydroxylation is dependent on an unidentified enzyme (Fig. 1a).

As part of ongoing high-throughput efforts, the New York Structural Genomics Research Consortium (NYSGRC) determined the structure of *Escherichia coli* CmoA, which revealed unexpected electron-density features at the predicted SAM binding site. When SAM was modelled, residual electron density suggestive of a carboxylate group was observed adjacent to the *S*-methyl group. Refinement of this structure at 1.50 ångström (Å) supports the idea that there is a covalent link between the *S*-methyl group and the putative carboxylate, consistent with the metabolite carboxy-SAM (Cx-SAM; (2S)-4-[{[(2S,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-3,4-dihydroxy-tetrahydrofuran-2-yl]methyl}(carboxylatomethyl)sulfonio]-2-ammoniobutanoate), which was unknown previously (Fig. 2 and Supplementary Fig. 1). Liquid chromatography–mass spectrometry (LC–MS) analysis of purified CmoA, confirmed Cx-SAM as the CmoA-bound ligand that persisted through purification and crystallization (Fig. 3 and Supplementary Fig. 2).

The overall structure of *E. coli* CmoA is similar to that previously reported for the *Haemophilus influenzae* orthologue (67% sequence identity; PDB entry 1IM8; root-mean-square deviation of 0.51Å for 222 equivalent C_α atoms)[7]. Retrospective refinement of the *H. influenza* CmoA structure revealed electron-density features consistent with Cx-SAM and contacts similar to those observed in the *E. coli* enzyme. These observations support the existence of conserved pathways involving Cx-SAM in the Gram-negative bacteria.

The *S*-carboxymethyl group of Cx-SAM in the *E. coli* CmoA catalytic site forms a bidentate polar interaction with the side-chain guanidinium of Arg 199, which is invariant among CmoA orthologues. The 2′- and 3′-hydroxyl groups of Cx-SAM form hydrogen bonds with the side chain of Asp 89; the equivalent residue in all other SAM-dependent methyltransferases is aspartate or glutamate. Other highly conserved residues among CmoA orthologues are contributed from helices α1, α6 and α7 (Supplementary Fig. 3), which seem to be crucial for substrate binding and are not present in other members of the SDMT superfamily. The Cx-SAM binding pocket is mainly hydrophobic, with no functionality capable of deprotonating the *S*-methyl group of SAM closer than 4.6 Å. Adjacent to the Cx-SAM binding site is a partially hydrophobic cavity, which is likely to be the binding site for an additional ligand or substrate (see below).

As there are data implicating chorismate, or a related metabolite, in wobble uridine oxyacetylation in Gram-negative bacteria[5,6], chorismate was examined as the potential carboxyl donor. Chorismate supported the CmoA-mediated formation of Cx-SAM, as demonstrated using LC–MS (Fig. 3b) and MS/MS (Supplementary Fig. 4). In addition to Cx-SAM, phenylpyruvate was produced with similar kinetics (Fig. 3c and Supplementary Fig. 5). The lack of a facile pathway for the direct conversion of chorismate to phenylpyruvate suggested that chorismate undergoes a rearrangement before CmoA-catalysed Cx-SAM formation. Among the several biologically characterized products of chorismate, only prephenate possesses a scaffold consistent with the observed LC–MS and MS/MS fragmentation data.

[1]Department of Biochemistry, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [2]Department of Pathology, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [3]Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94158, USA. [4]Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, San Francisco, California 94158, USA. [5]Department of Biology, Johns Hopkins University, Baltimore, Maryland 21218, USA. [6]Department of Physiology & Biophysics, Albert Einstein College of Medicine, Bronx, New York 10461, USA.

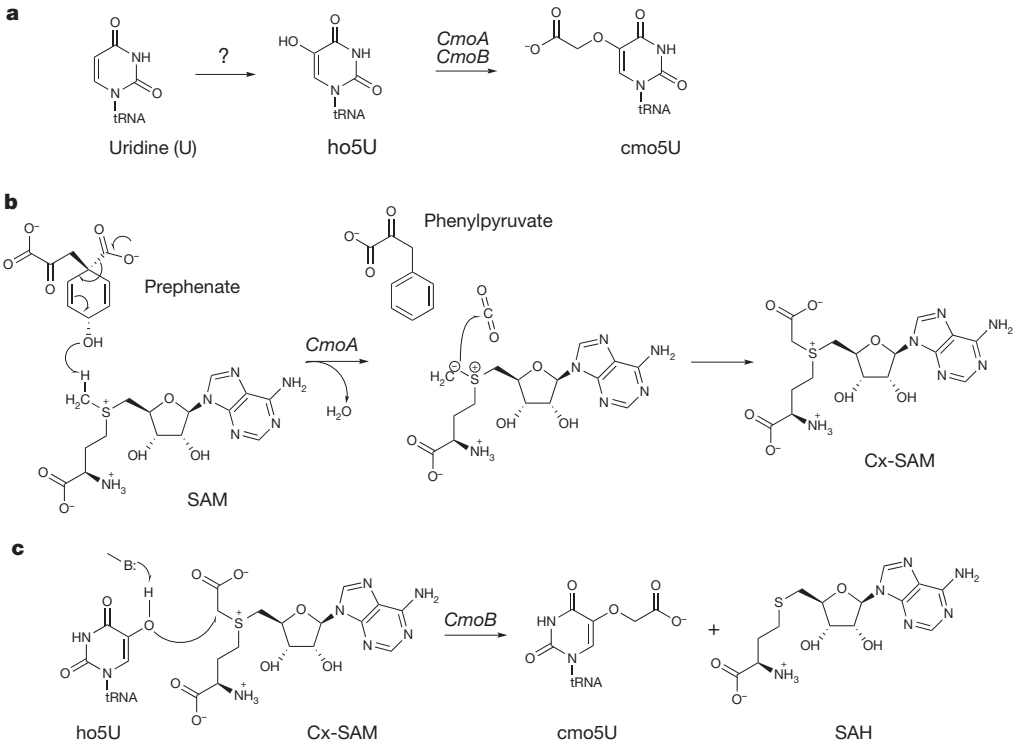


Figure 1 | Proposed chemical mechanism for the biosynthesis of cmo5U. **a**, Previously identified biosynthetic pathway for cmo5U at wobble uridines. First, the wobble uridine is converted to ho5U by an unknown mechanism, and this is followed by the action of CmoA and CmoB. **b**, Mechanism for CmoA-catalysed Cx-SAM formation from SAM and prephenate. **c**, Mechanism for CmoB-catalysed formation of cmo5U from ho5U and Cx-SAM.

When the *in vitro* production of Cx-SAM was investigated, prephenate was found to be a more efficient substrate than chorismate (Fig. 3d) and the lag in Cx-SAM production observed with chorismate was absent. This behaviour is consistent with the slow non-enzymatic conversion of chorismate to prephenate, which is used by CmoA for Cx-SAM formation (Supplementary Fig. 6). Furthermore, production of phenylpyruvate from prephenate was confirmed by NMR and LC–MS of the *in vitro* assay solution (Supplementary Figs 7 and 8). These activities are CmoA-specific as CmoB does not show any Cx-SAM synthase activity (data not shown). We propose that prephenate is the biologically relevant substrate and the source of the carboxylate in the CmoA-catalysed reaction.

These results are consistent with the CmoA-catalysed decarboxylation and concomitant loss of hydroxide from prephenate to produce phenylpyruvate, water and carbon dioxide, which is the source of the carboxylate functionality in Cx-SAM (Fig. 1b). In support of this hypothesis, uniformly ^{13}C-labelled chorismate ([U-^{13}C] chorismate) transferred ^{13}C-carbon dioxide to the product, Cx-SAM (Supplementary Fig. 9), demonstrating that prephenate is the carboxyl donor in the CmoA-catalysed formation of Cx-SAM and validating the overall reaction proposed for CmoA. To our knowledge, this is the first example of prephenate serving as the carboxyl group donor in a carboxytransfer reaction.

There is precedence for decarboxylation of prephenate, as prephenate dehydratase (PDT) catalyses the decarboxylation of prephenate with concomitant loss of hydroxide to generate phenylpyruvate in a fashion similar to that suggested for CmoA[8]. In the *Methanocaldococcus jannaschii* PDT-catalysed reaction, elimination of the hydroxyl group as water is facilitated by the participation of Thr 172 as a general acid to protonate the leaving hydroxide group. The threonine side chain is not sufficiently acidic to protonate the prephenate hydroxyl directly (that is, the conjugate acid of the prephenate hydroxyl is expected to behave in a similar manner to an alcohol, with a pK_a of approximately −2, whereas the pK_a of threonine is approximately 16 in water); thus, other mechanisms must be operative as the major driving force for this reaction. It was proposed previously that geometric distortion drives decarboxylation and that the favourable energetics associated with aromatization reduce the bond order of the hydroxyl, shifting its reactivity towards that of hydroxide and enabling efficient general acid catalysis[8]. Similar considerations are relevant to the CmoA-catalysed reaction.

Computational docking of prephenate in the CmoA catalytic site suggests a pose in which the carboxylate and hydroxyl groups of prephenate sandwich the *S*-methyl group of SAM, consistent with the observed transcarboxylation reaction (Fig. 2 and Supplementary Fig. 10). The importance of carbon–oxygen (CH···O) hydrogen bonds in the recognition and presentation of the *S*-methyl group in canonical

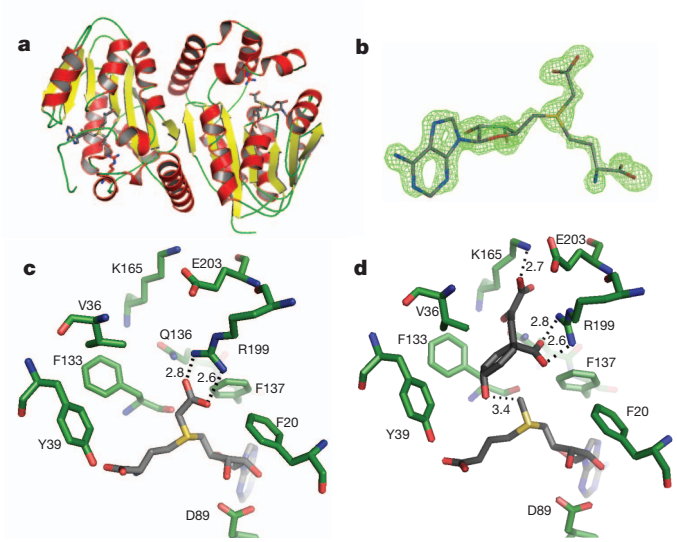


Figure 2 | Structure of the CmoA–Cx-SAM complex. **a**, Overall dimeric structure of *E. coli* CmoA, with α-helices, β-sheets and loops shown in red, yellow and green, respectively. Cx-SAM and Arg 199 are represented as sticks. **b**, $F_o - F_c$ difference Fourier synthesis, calculated at 1.5 Å resolution with the ligand omitted, contoured at 5σ around the modelled Cx-SAM ligand. **c**, Catalytic site of CmoA. Protein carbon atoms are shown in green and Cx-SAM carbon atoms in grey. Oxygen and nitrogen atoms are shown in red and blue, respectively. Ionic interactions between Cx-SAM and the side chain of Arg 199 are depicted as dashed lines (distances in Å). **d**, Computationally predicted pose of prephenate in the CmoA catalytic site.

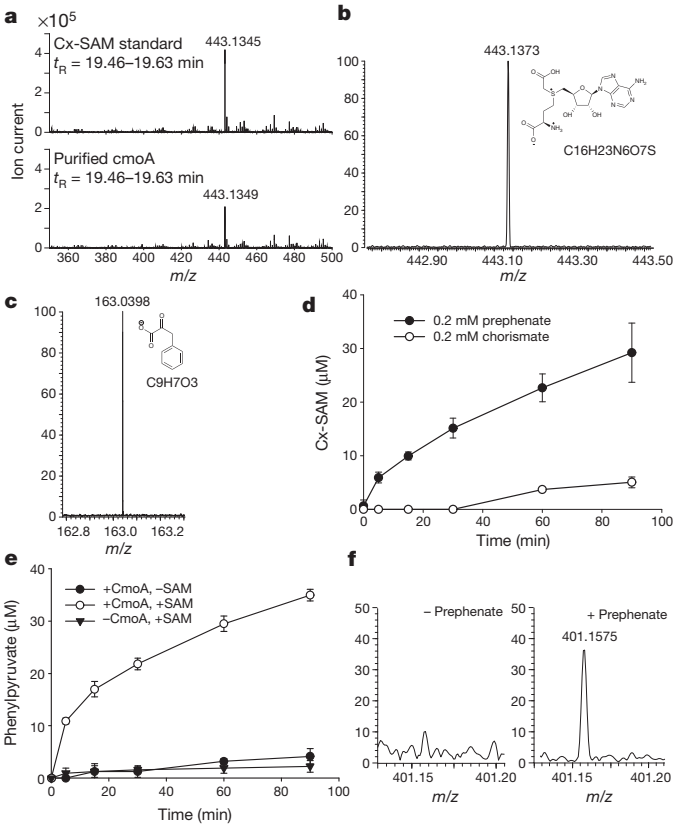


Figure 3 | Identification of low-molecular-weight compounds associated with CmoA-mediated Cx-SAM production. **a**, Electrospray time-of-flight (ESI-TOF) mass spectra of a Cx-SAM standard (top) and the low-molecular-weight compound co-purifying with CmoA (bottom). Peak mass-to-charge ratio (m/z) were 443.1345 and 443.1349 (errors, −0.9 and +2.3 p.p.m.) for Cx-SAM standard and the compound that co-purified with recombinant CmoA, respectively. **b**, Detection of Cx-SAM in an *in vitro* assay containing SAM, chorismate and CmoA. **c**, Detection of phenylpyruvate ($C_9H_7O_3$) formation in the assay mixture by mass spectrometry in negative mode (m/z = 163.0398 observed, 163.0395 calculated). **d**, Time course of Cx-SAM production in an *in vitro* assay of CmoA. The assay solution contained 20 mM sodium phosphate, pH 6.8, 0.2 mM [^{14}C-methyl]-SAM, 0.2 mM prephenate or chorismate, and 2 μM CmoA. Error bars represent the s.d. of three data sets. **e**, Time course of the phenylpyruvate formation from prephenate. The assay mixture contained 20 mM sodium phosphate, pH 6.8, 0.2 mM prephenate, 0.2 mM SAM (open circles and inverted triangles) and 2 μM CmoA (open circles and filled circles). Error bars represent the s.d. of three data sets. **f**, Solvent proton exchange of [$^2H_3$-methyl]-SAM catalysed by CmoA. The sample contained 10 mM Tris, pH 8.0, 0.5 mM [$^2H_3$-methyl]-SAM and 10 μM CmoA, with or without 0.5 mM prephenate. The reaction was carried out at room temperature (20 to 25 °C) for 4 h. In the presence of prephenate, doubly deuterated SAM (calculated m/z = 401.1576) was observed. t_R, retention time.

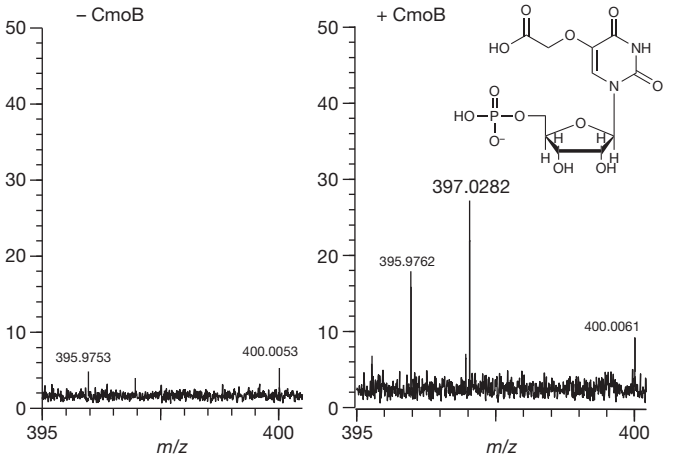


Figure 4 | *In vitro* assay of CmoB-catalysed carboxymethyltransfer activity. Total RNA extracted from CmoB mutant cells was used as a substrate, and Cx-SAM was generated *in situ* by the action of CmoA on prephenate and SAM. RNA was digested with P1 nuclease to 5′-nucleotide monophosphates and then mass-spectrometry analysis was carried out. Left, no CmoB was added; right, addition of purified *E. coli* CmoB resulted in the detection of cmoUMP ($C_{11}H_{14}O_{12}N_2P$, m/z = 397.0282 observed, 397.0284 calculated) in negative ion mode.

S_N2-based SAM-dependent methyltransfer reactions was highlighted recently[9]. In the prephenate-bound model of CmoA (Fig. 2 and Supplementary Fig. 10), the hydroxyl oxygen of prephenate and the *S*-methyl carbon of SAM form a potential 3.4-Å CH···O hydrogen bond. Notably, the prephenate hydroxyl is poorly positioned for an in-line S_N2 attack on the *S*-methyl group, consistent with the lack of prephenate methylation. Instead, this arrangement suggests a substrate-assisted mechanism in which the departing prephenate OH group abstracts a proton from the *S*-methyl group of SAM, generating water and the nucleophilic ylide (stabilized carbanion), which is carboxylated by the carbon dioxide (Fig. 1b). Importantly, the pK_a of the trimethylsulphonium cation (a model of the *S*-methyl group in SAM) has been reported as 18.9 and 18.2 in water and dimethylsulphoxide (DMSO)[10,11], respectively. These values are similar to that of the general acid (Thr 172) in the PDT-catalysed reaction and are consistent with elimination of hydroxide from prephenate as water and formation of the reactive ylide intermediate.

We sought direct evidence for the formation of the sulphonium ylide, given its importance in the mechanism. Presentation of CmoA with triply deuterated SAM ([$^2H_3$-methyl]-SAM) and prephenate results in formation of doubly deuterated SAM, consistent with formation of the ylide intermediate by deuteron abstraction and regeneration of SAM by protonation (Fig. 3f). In the absence of prephenate, only triply deuterated SAM was observed. In the absence of CmoA, no exchange was observed in mixtures of [$^2H_3$-methyl]-SAM and prephenate. Partitioning of the ylide between product formation (Cx-SAM) and its return to substrate (SAM) was quantified by determining the ratio between solvent-exchanged SAM and Cx-SAM: 92.0:8.0 and 97.3:2.7 at pH 7.3 and 8.5, respectively, providing strong evidence that the postulated ylide intermediate is on the reaction coordinate for Cx-SAM formation.

The most common biological fate of the SAM *S*-methyl group is intermolecular transfer catalysed by methyltransferases, including epigenetic marking of DNA and histone targets[12,13], and a wide range of small molecule transformations[14,15]. We proposed that Cx-SAM is used in a CmoB-catalysed transcarboxymethylation reaction in the biosynthesis of cmo5U. Total RNA[16] and purified tRNAs[2] from cmoB-deficient cells (that is, ho5U-containing RNA) were used as substrates in *in vitro* assays with prephenate, SAM and CmoA, with or without recombinant CmoB. After the transfer reaction, RNAs were treated with P1 nuclease and the resulting 5′-nucleotide monophosphates analysed by mass spectrometry (Fig. 4). CmoB catalysed carboxymethyl transfer from *in situ*-generated Cx-SAM to ho5U-containing RNAs, as a species with mass corresponding to oxyacetyl-5-uridine-5′-monophosphate (cmo5UMP) was clearly detected. CmoA alone does not exhibit carboxymethyltransferase activity. Therefore, we conclude that Cx-SAM is the substrate for the CmoB-dependent transcarboxymethylation of ho5U-containing tRNAs (Fig. 1c).

In most Gram-negative species, *cmoA* and *cmoB* are co-conserved and immediately adjacent to each other in the genome, supporting the demonstrated functional relationship. Most importantly, the unique *in vitro* activities assigned to CmoA and CmoB are fully consistent with all reported genetics findings relevant to cmo5U modification[2,5,6]. Our own genetics and mutagenesis studies add further support, as plasmid-based expression of wild-type CmoA restores production of cmo5U in CmoA-deficient (*ΔcmoA*) *E. coli*, whereas the Asp89Leu mutant

lacking *in vitro* biochemical activity failed to complement *in vivo* (Supplementary Fig. 11). Finally, *in vivo*- and *in vitro*-generated Cx-SAM-bound CmoA exhibited comparable behaviour in the carboxymethylation of ho5U-containing RNAs, supporting the relevance of our *in vitro* studies (see Supplementary Fig. 12).

Elaboration of the *S*-methyl group of SAM with a variety of functional groups has been pursued by chemical biologists to support studies including the generation of modified nucleic acid[17,18] and protein[19] substrates, as well as genome-wide assignments of protein methyltransferase targets[19]. It is notable that Cx-SAM represents the first SAM derivative demonstrated to have been shaped by evolution for a biologically meaningful function. The electrophilic properties of the SAM sulphonium centre have been exploited to realize expanded functional diversity, as exemplified by Cx-SAM, through two unique activities within the SDMT superfamily (Supplementary Fig. 13a). Sequence analysis supports the existence of CmoA orthologues throughout the Gram-negative proteobacteria, as well as in the *Verrucomicrobia* and some *Cyanobacteria* (Supplementary Fig. 13b). It remains to be discovered how widespread these mechanisms are and whether additional biologically relevant SAM-analogues exist.

The fortuitous identification of a bound ligand in a crystal structure (such as Cx-SAM in CmoA) is not unusual. We estimate that approximately 3 to 5% of all structures determined by the NYSGRC contain organic ligands derived from the expression host (typically *E. coli*) that persisted through purification and crystallization. Frequently this can be anticipated; for example, finding NAD or NADH bound to targets annotated as oxidoreductases, or pyridoxal phosphate in annotated aminotransferases. However, unanticipated ligands are also identified, including nucleotides, amino acids, carbohydrates and lipids bound to proteins from a range of bacterial species, providing clues to catalytic activity and biological function (see examples in Supplementary Fig. 14).

In summary, direct structural observation identified the novel metabolite Cx-SAM, leading to the discovery of unique Cx-SAM synthase and carboxymethyltransferase activities involved in tRNA wobble base modification. These findings reveal new functional diversity in the SDMT superfamily, expand the metabolic and biological contributions for SAM-based biochemistry, and presage the discovery of new metabolites and biological processes. This work highlights the power of structural genomics approaches for the discovery of new metabolites, pathways and biology.

METHODS SUMMARY

CmoA and CmoB were expressed in *E. coli* and purified by Ni-NTA (nickel-nitriloacetic acid) and size-exclusion chromatography. CmoA was crystallized by vapour diffusion and the structure was determined by molecular replacement. CmoA-catalysed formation of carboxyl-SAM was monitored using [^{14}C-methyl]-SAM and either chorismate or prephenate as the carboxyl donor. Phenylpyruvate formation was monitored spectroscopically at 320 nm and ylide formation was monitored by solvent isotope exchange of trideuterated-SAM ([$^{2}H_3$-methyl]-SAM). CmoB carboxymethyltransferase activity was assessed by the modification of RNAs purified from CmoB-deficient *E. coli*. Assay products were assigned by LC–MS, mass-spectrometry fragmentation and nuclear magnetic resonance (NMR).

Full Methods and any associated references are available in the online version of the paper.

Received 23 July 2012; accepted 12 April 2013.
Published online 15 May 2013.

1. Czerwoniec, A. *et al.* MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res.* **37,** D118–D121 (2009).
2. Nasvall, S. J., Chen, P. & Bjork, G. R. The modified wobble nucleoside uridine-5-oxyacetic acid in tRNAPro(cmo5UGG) promotes reading of all four proline codons *in vivo*. *RNA* **10,** 1662–1673 (2004).
3. Näsvall, S. J., Chen, P. & Bjork, G. R. The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. *RNA* **13,** 2151–2164 (2007).
4. Weixlbaumer, A. *et al.* Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nature Struct. Mol. Biol.* **14,** 498–502 (2007).
5. Björk, G. R. A novel link between the biosynthesis of aromatic amino acids and transfer RNA modification in *Escherichia coli*. *J. Mol. Biol.* **140,** 391–410 (1980).
6. Hagervall, T. G., Jonsson, Y. H., Edmonds, C. G., McCloskey, J. A. & Bjork, G. R. Chorismic acid, a key metabolite in modification of tRNA. *J. Bacteriol.* **172,** 252–259 (1990).
7. Lim, K. *et al.* Crystal structure of YecO from *Haemophilus influenzae* (HI0319) reveals a methyltransferase fold and a bound S-adenosylhomocysteine. *Proteins* **45,** 397–407 (2001).
8. Van Vleet, J., Kleeb, A., Kast, P., Hilvert, D. & Cleland, W. W. 13C isotope effect on the reaction catalyzed by prephenate dehydratase. *Biochim. Biophys. Acta* **1804,** 752–754 (2010).
9. Horowitz, S., Yesselman, J. D., Al-Hashimi, H. M. & Trievel, R. C. Direct evidence for methyl group coordination by carbon-oxygen hydrogen bonds in the lysine methyltransferase SET7/9. *J. Biol. Chem.* **286,** 18658–18663 (2011).
10. Crosby, J. & Stirling, C. J. M. Elimination and addition reactions. Part XIX. Elimination of phenoxide from β-substituted ethyl phenyl ethers: the nature of activation in 1,2-elimination. *J. Chemic. Soc. B* 671–679 (1970).
11. Bordwell, F. G. Equilibrium acidities in dimethyl sulfoxide solution. *Acc. Chem. Res.* **21,** 456–463 (1988).
12. Arrowsmith, C. H., Bountra, C., Fish, P. V., Lee, K. & Schapira, M. Epigenetic protein families: a new frontier for drug discovery. *Nature Rev. Drug Discov.* **11,** 384–400 (2012).
13. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Rev. Genet.* **10,** 295–304 (2009).
14. Luka, Z., Mudd, S. H. & Wagner, C. Glycine *N*-methyltransferase and regulation of *S*-adenosylmethionine levels. *J. Biol. Chem.* **284,** 22507–22511 (2009).
15. Vévodová, J. *et al.* Structure/function studies on a *S*-adenosyl-L-methionine-dependent uroporphyrinogen III C methyltransferase (SUMT), a key regulatory enzyme of tetrapyrrole biosynthesis. *J. Mol. Biol.* **344,** 419–433 (2004).
16. Kowtoniuk, W. E., Shen, Y., Heemstra, J. M., Agarwal, I. & Liu, D. R. A chemical screen for biological small molecule-RNA conjugates reveals CoA-linked RNA. *Proc. Natl Acad. Sci. USA* **106,** 7768–7773 (2009).
17. Dalhoff, C., Lukinavicius, G., Klimasauskas, S. & Weinhold, E. Direct transfer of extended groups from synthetic cofactors by DNA methyltransferases. *Nature Chem. Biol.* **2,** 31–32 (2006).
18. Dalhoff, C., Lukinavicius, G., Klimasauskas, S. & Weinhold, E. Synthesis of *S*-adenosyl-L-methionine analogs and their use for sequence-specific transalkylation of DNA by methyltransferases. *Nature Protocols* **1,** 1879–1886 (2006).
19. Binda, O. *et al.* A chemical method for labeling lysine methyltransferase substrates. *ChemBioChem* **12,** 330–334 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Parker and C. T. Walsh for providing the *Aerobacter aerogenes 62-1* strain. We are indebted to V. Schramm and J. Gerlt for critical discussions and reading of the manuscript. This work was supported by US National Institutes of Health grants GM094662 (to S.C.A.), GM093342 (to S.C.A., M.P.J. and P.C.B.) and the Albert Einstein Cancer Center. This publication was made possible by the Center for Synchrotron Biosciences grant P30-EB-009998 from the National Institute of Biomedical Imaging and Bioengineering (NIBIB). Use of the National Synchrotron Light Source, Brookhaven National Laboratory, was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract no. DE-AC02-98CH10886.

Author Contributions J.K. carried out cloning, protein purification, crystallography, and functional assays. H.X. performed mass-spectrometry analysis of the *in vitro* assay. Y.-S.L. carried out LC–MS analysis of the CmoA-bound ligand and chemical synthesis of Cx-SAM. X.T. performed the NMR experiments. N.F.A.-O. carried out thermal denaturation studies. C.K. and M.P.J. performed computational modelling. S.B. and P.C.B. carried out the bioinformatics analysis. J.B.B. and Y.P. assisted in crystallographic validation and analysed crystallographic ligand-binding results. J.K. and S.C.A. designed the study, analysed the data and wrote the manuscript.

Author Information Atomic coordinates and structure factors for the reported crystal structure are deposited in the Protein Data Bank under the accession code 4GEK. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.K. (jungwook.kim@einstein.yu.edu) or S.C.A. (steve.almo@einstein.yu.edu).

METHODS

Cloning and protein purification. The *cmoA* gene was amplified from genomic DNA of *E. coli* BL21 by polymerase chain reaction (PCR), cloned into LIC-pET46a (Novagen) and verified by the DNA sequence analysis (Genewiz). *E. coli* BL21 (DE3) cells (Invitrogen) were transformed with vectors harbouring the *cmoA* gene, grown in lysogeny broth containing 100 mg ml^{-1} ampicillin at 37 °C and induced with 0.5 mM isopropylthiogalactoside (IPTG) at an $D_{600\,nm}$ of approximately 1. Cells were incubated overnight at 25 °C and collected by centrifugation. Cell pellets were resuspended with Bugbuster (Novagen) at room temperature (20 to 25 °C) for 30 min, the lysates were centrifuged at 41,400*g* for 30 min, and the supernatants were applied to Ni-agarose (Qiagene) columns pre-equilibrated with buffer A (20 mM HEPES, pH 7.5, and 150 mM KCl). The recombinant protein was eluted with 150 mM imidazole in buffer A and purified further by size-exclusion chromatography on a HiLoad Superdex 200 column (GE) equilibrated with buffer A. Final purity was over 95% as verified by SDS–PAGE (SDS–polyacrylamide gel electrophoresis) analysis. For the measurement of $D_{280\,nm}$ of the purified CmoA, the enzyme was denatured to separate SAM or Cx-SAM from the protein using multiple rounds of mixing with 8 M guanidine chloride solution followed by spin filtration. An extinction coefficient of $\varepsilon_{280} = 18.7\,cm^{-1}\,mM^{-1}$ was used to calculate the yield of the nucleoside-free CmoA as calculated from the amino acid sequence. The *cmoB* gene was amplified from genomic DNA of *E. coli* BL21 by PCR and cloned into LIC-pET30a (Novagen). The purification of *E. coli* CmoB was identical to that of CmoA, except for the use of kanamycin as the selectable marker. In addition, the affinity tag was removed by thrombin (Novagen) cleavage after elution of the recombinant CmoB from the Ni-agarose resin. The yield was quantified using an extinction coefficient of $\varepsilon_{280} = 72.5cm^{-1}\,mM^{-1}$, as calculated from the amino acid sequence.

The Asp89Leu mutant of CmoA was generated by QuickChange (Stratagene) with primers 5′-TTGCAAAATTATTGCCATCCTCAACTCCCCGGCGATGATT-3′ and 5′-AATCATCGCCGGGGAGTTGAGGATGGCAATAATTTTGCAA-3′, and the plasmid of the wild-type CmoA as the template for PCR. Purification of the Asp89Leu mutant was similar to that of the wild type described above; some purifications included the addition of anion exchange and gel-filtration chromatography steps. For anion exchange separation, a MonoQ column (GE) was equilibrated with Buffer A (20 mM Tris-HCl, pH 8.5, and 150 mM KCl) and a 1-ml sample was loaded. A linear gradient of Buffer B (20 mM Tris-HCl, pH 8.5, and 1 M NaCl) was used to elute bound protein. Wild-type CmoA eluted as a single peak on MonoQ, whereas the Asp89Leu mutant showed two peaks. Eluted proteins were analysed by SDS–PAGE, pooled, concentrated and loaded on a Superdex75 column (GE) equilibrated with Buffer A.

Crystallization and structural determination of CmoA. Purified CmoA was crystallized using the method of sitting drop vapour diffusion at 21 °C by mixing 1 µl of the protein with 1 µl of reservoir solution (0.2 M Li_2SO_4, 0.1 M Bis-Tris:HCl (HCl used to adjust the pH of the buffer), pH 5.5, and 25% PEG3350) and equilibrating over 0.1 ml of reservoir solution. X-ray data were collected on an ADSC QUANTUM 315 CCD detector at the National Synchotron Light Source (NSLS) beam line X29A and processed with HKL3000 (ref. 20). Diffraction data from CmoA crystals were collected at 100 K, and at a wavelength (λ) of 0.9790 Å, which were consistent with space group, $P2_12_12$ (a = 65.32 Å, b = 78.68 Å, c = 92.37 Å), with two molecules per asymmetric unit. Molecular replacement was carried out using the structure of *H. influenzae* YecO (PDB code 1IM8) as the search model with MOLREP[21]. Subsequent model building and refinement were performed with Coot[22] and REFMAC5[23]. The final model was refined to 1.50 Å with *R* values of $R_{work} = 0.17$ and $R_{free} = 0.20$ (Supplementary Table 1). All residues are in energetically allowed regions of the Ramachandran plot.

Time-course assay of Cx-SAM production. The time-dependent formation of carboxyl-SAM was monitored using [^{14}C-methyl]-SAM with either prephenate or chorismate. The assay mixture contains 20 mM sodium phosphate, pH 6.8, 0.2 mM [^{14}C-methyl]-SAM (Perkin Elmer), and either 0.2 mM chorismate or prephenate. The assay was initiated by adding purified 2 µM CmoA to the assay mixture, with a total volume of 20 µl. A 2-µl aliquot was withdrawn periodically and mixed with an equal volume of 0.1 M HCl to quench the reaction, then 1 µl was spotted onto a thin layer chromatography (TLC) plate. The TLC was developed with buffer composed of 79% ammonium sulphate, 19% iso-propanol and 2% sodium acetate, pH 6.0. The plate was air-dried and exposed to a phosphor screen imager (GE) for 2 days. The image plate was scanned using a Molecular Dynamics Storm 860 PhosphorImager System with ImageQuant software.

Time-course assay of phenylpyruvate production. Phenylpyruvate formation was monitored at a wavelength of 320 nm as described previously[24]. The assay mixture contained 20 mM sodium phosphate, pH 6.8, and 0.2 mM prephenate, with or without 0.2 mM SAM in a total volume of 0.5 ml. The reaction was initiated by adding 2 µM CmoA to the assay solution. A 70-µl aliquot of the reaction mixture was withdrawn periodically and added to 30 µl of 5 M NaOH. The absorbance at a wavelength of 320 nm was measured and the net $D_{320\,nm}$ was recorded by subtracting residual absorbance arising from contaminating phenylpyruvate. The net absorbance was converted to the concentration of phenylpyruvate using a standard curve prepared with commercially obtained phenylpyruvate (Sigma-Aldrich). Non-enzymatic turnover of prephenate to phenylpyruvate was measured with a sample composed of 20 mM sodium phosphate, pH 6.8, 0.2 mM prephenate, and 0.2 mM SAM in 0.5 ml, without the addition of the enzyme.

Verification of Cx-SAM co-purified with recombinant CmoA by LC–MS. A 10-µl recombinant protein solution (10 mg ml^{-1}) was diluted with 10 µl water and then with 190 µl of methanol. The mixture was centrifuged at room temperature for 10 min (16,000g), and the supernatant was used for the analysis. For each injection, an 80-µl aliquot was subjected to LC–MS analysis (Agilent 1200 HPLC coupled with Agilent 6210 AccurateMass electronspray mass spectrometer; ESI positive ion mode detection, 4 GHz, *m/z* range from 50 to 1,200; Phenomenex Luna NH2 column, bead size of 5 µm, pore size of 100 Å, 150 × 2 mm) using a gradient system described in the literature[25]. Data were analysed using the Agilent Mass Hunter software package.

Mass-spectroscopy analysis of CmoA assay. Prephenate (10 mM; or 10 mM chorismate initially) was incubated with 10 mM SAM and 10 µM CmoA in a total of 0.5 ml solution at room temperature overnight. An aliquot of 10 µl of the reaction mixture was mixed with 100 µl methanol, which was then infused into a 12T Agilent IonSpec FT-ICR-MS (Agilent Technologies). Cx-SAM ($m/z = 443.1373$) was monitored in positive mode, and phenylpyruvate ($m/z = 163.0404$) was monitored in negative mode. The Agilent 12T QFT-ICR routinely provides better than 5 p.p.m. mass accuracy with external calibration.

Solvent isotope exchange of deuterated SAM. The assay solution was prepared by mixing 10 mM Tris, pH 8.0, 0.5 mM [$^2H_3$-methyl] SAM (CDNisotope), 0.5 mM prephenate and 10 µM CmoA in 0.5 ml solution. The reaction was incubated for 4 h at room temperature and was quenched by filtering the enzyme with a spin column (molecular-weight cut-off (MWCO), 10 kilodaltons (kD)). The sample was then analysed by mass spectrometry as described above. To examine whether solvent proton exchange at *S*-methyl of SAM is prephenate-dependent in manner, a sample without prephenate was prepared and analysed in an identical fashion.

Assay of CmoB reaction. Carboxymethyltransfer activity of CmoB was examined with a solution containing 50 mM Tris, pH 8.0, 4 mM $MgCl_2$, 1 mM prephenate, 1 mM SAM, 1 µM CmoA and total RNA extracted from *cmoB*-mutant *E. coli* cells as described before[16], or purified tRNAs[2]; the total volume of the assay was 50 µl. The reaction was initiated by adding 6 µM CmoB and incubated at the room temperature for 2 h. One unit of P1 nuclease (US Biological) was added to the assay solution and incubated at 65 °C for 1 h to convert polynucleotides into 5′-nucleotide monophosphates. The P1 nuclease-treated sample was mixed with 100 µl methanol and vortexed before centrifugation at 16,000*g* for 2 min. An aliquot of supernatant was injected to a 12T Agilent IonSpec FT-ICR-MS and analysed in negative mode.

Computational ligand docking. To create a model of the substrate prephenate bound to CmoA, we first removed the carboxylate group from the product Cx-SAM to create SAM in the CmoA catalytic site. The resulting complex was subjected to a protein preparation protocol, during which hydrogen atoms were added, protonation states of His residues were examined and adjusted if necessary, side-chains of Thr, Tyr and Asn residues were optimized for hydrogen bonding interactions, and the entire structure was finally energy-minimized such that heavy atom positions remained within 0.3 Å of the starting coordinates.

Initial attempts to dock the substrate prephenate to this resulting model failed owing to inadequate space for the ligand. We proposed that charged residues in the binding site required conformational changes to accommodate the ligand. Specifically, in the Cx-SAM-bound structure, Arg 199 formed a salt-bridge interaction with the carboxylate group of Cx-SAM (Fig. 2c), and other charged residues in the active site such as Lys 165 and Glu 203 either pointed towards the solvent or blocked portions of the active site. Therefore, we used an induced-fit docking procedure, in which side chains of residues that are within 5 Å of the docked prephenate pose were treated as conformationally flexible[26]. Induced-fit docking uses a combination of a molecular-mechanics energy function and an empirical scoring function-based energy to rank the ligand poses. We re-ranked the induced fit docking poses using a molecular-mechanics-based energy function that has been used successfully in many applications of metabolite docking. The lowest energy binding pose identified using this technique is shown in Fig. 2d.

Thermal stability of wild-type and Asp89Leu mutant CmoA. The fluorescence-monitored thermal denaturation of wild-type and mutant CmoA was carried out using a 7900HT RT–PCR system (Applied Biosystems). In brief, 20 µl of each protein at 10 µM concentration was mixed with 0.5 µl of 200× Sypro orange solution and pipetted into separate wells of a 384-well PCR plate. After centrifugation to remove air bubbles, the plate is loaded into the PCR machine and the temperature ramped from 20 to 99 °C, in 1-°C increments with a dwell time of 6 s.

The negative first derivative of the fluorescence change (−dRFU/dT) for each protein is plotted against temperature, and the melting temperature is defined as the minimum in the −dRFU/dT curve. The wild type from MonoQ exhibited a melting temperature (T_m) of 55.6 ± 0.1 °C, whereas the two Asp89Leu mutant fractions had T_m that were 2 to 3 °C lower. The behaviour of the wild-type and mutant species show that they are folded under the conditions (that is, temperature) used in *in vitro* activity assays; the lack of activity exhibited by the Asp89Leu mutant is thus the consequence of a catalytic defect and not due to issues related to thermodynamic stability.

Chemical synthesis of Cx-SAM. *S*-Adenosyl-L-homocysteine (SAH; 3 mg) was dissolved in 0.5 ml of 150 mM ammonium bicarbonate. To this solution, 2-iodoacetatic acid (100 mg) was added. The mixture was incubated at 37 °C for 12 h with constant agitation. The progress of the reaction was monitored by TLC (SiO_2) using a solvent system composed of methanol:aqueous 1.5 N ammonium bicarbonate (10:1 vol/vol). Retardation factor (R_f) = 0.6 and 0.3 for SAH and Cx-SAM, respectively. After the reaction was completed, 12 ml methanol was added and the mixture was incubated at 4 °C overnight. Precipitates were collected by centrifugation at 4 °C (2,000*g* for 30 min), washed twice with ice-cold methanol and dissolved in 0.10 ml deionized water. The product was purified using hydrophilic interaction chromatography (HILIC) as described above. Concentration of Cx-SAM was determined spectroscopically, assuming an extinction coefficient of SAM ($\varepsilon_{260} = 15.4\,cm^{-1}\,mM^{-1}$).

Assay of non-enzymatic formation of prephenate from chorismate. The rate of conversion from chorismate to prephenate in the absence of CmoA was measured *in vitro*[24]. The assay solution contained 10 mM sodium phosphate (pH 6.8), 0.2 mM SAM, and 0.2 mM chorismate, in 0.5 ml. An aliquot of 80 µl was withdrawn periodically and added to 5 µL of 4.5 M HCl. The mixture was then incubated at 37 °C for 15 min and combined with 15 µl of 12 M NaOH before the absorbance at 320 nm was measured to determine phenylpyruvate.

NMR analysis of the CmoA assay mixture. To define the nature of prephenate-derived product subsequent to donation of the carboxylate, the *in vitro* assay was scaled up using 0.5 mM prephenate, 0.5 mM SAM, and 10 µM CmoA, in a total volume of 10 ml, to maximize the yield of products for NMR analysis. The reaction was incubated at room temperature for 8 h, and the enzyme was filtered using a spin column (MWCO, 10 kD) before lyophilization. The lyophilized sample was dissolved in 0.6 ml D_2O (Cambridge Isotope Laboratory) and ^{1}H-resonance data was collected with a Bruker DRX-300 NMR spectrometer.

Partitioning of the ylide intermediate. The pH-dependent ylide partitioning assay was carried out with 0.2 mM prephenate, 0.05 mM [$^2H_3$-methyl] SAM in either 10 mM ammonium acetate, pH 7.3, or 10 mM ammonium bicarbonate, pH 8.5. The assay was initiated by adding 2 µM CmoA, with incubation at room temperature for 2 h before analysis with mass spectrometry as described below. The total amount of [$^2H_3$-methyl]- and [$^2H_2{}^1$H-methyl]-SAM remaining after the reaction was determined by adding a known amount of unlabelled SAM to the assay mixture as an internal standard. The concentration of Cx-SAM was determined by subtracting the remaining SAM after the reaction from the initial quantity added. The amount of solvent exchanged ([$^2H_2{}^1$H-methyl]) and non-exchanged ([$^2H_3$-methyl]) SAM were calculated from the relative amplitude of corresponding mass-spectrometry peaks and the total SAM concentration. Based on the absolute concentrations of [$^2H_2{}^1$H-methyl]-SAM and Cx-SAM, the partitioning of the ylide intermediate back to reactant and forwards to product was calculated (for example, partitioning back to SAM is calculated as ([$^2H_2{}^1$H-methyl]-SAM)/([$^2H_2{}^1$H-methyl]-SAM + Cx-SAM)).

Verification of enzymatic formation of phenylpyruvate using LC–MS. A Shimadzu HPLC, with two LC-20AD pumps, was used to generate a gradient with 50 µl min^{-1} flow rate. Solvent A was 5% acetonitrile in H_2O and 0.1% formic acid, and solvent B consisted of 95% acetonitrile in H_2O and 0.1% formic acid. The assay sample (50 µl), which was used for NMR analysis above, was loaded onto a 1.0 × 50-mm C18 column (Phenomenex). After desalting with solvent B for 5 min, bound phenylpyruvic acid was eluted with a 30-min gradient composed of 5% to 95% solvent B. The effluent was delivered directly into the 12T QFT-ICR-MS (Agilent Technologies) for mass analysis.

Preparation of [U-^{13}C] chorismate. Production and purification of chorismate using *Aerobacter aerogenes 62-1* was carried out using the methods developed previously[27,28]. Medium A (for cell growth) and medium B (for chorismate synthesis) were prepared as described previously, except that glucose was omitted from medium B. *A. aerogenes 62-1* was a gift of J. Parker and C. T. Walsh. Overnight culture (1 ml) of *A. aerogenes 62-1* was added to 50 ml medium A, and incubated at 30 °C until $D_{600\ nm}$ reached approximately 1. Cells were pelleted at 3,000*g* and washed with 25 ml medium B, which did not contain glucose. After pelleting once again, cells were resuspended in 25 ml medium B with 0.2 g of [U-^{13}C] glucose (purchased from Cambridge Isotope Laboratory). Cells were grown at 30 °C for 15 h for the production of labelled chorismate, collected by centrifugation and discarded. The supernatant was filtered and loaded on a Hypercarb HPLC column (10 × 100 mm) equilibrated in 10 mM ammonium acetate, pH 9.9. Chorismate was eluted with a linear gradient of acetonitrile and identified by monitoring $D_{275\ nm}$ of each fraction. ^{13}C-chorismate was confirmed by mass spectrometry ($^{13}C_{10}H_{10}O_6Na$, m/z = 259.0707 observed, 259.0711 calculated). Pooled chorismate was lyophilized, resuspended in water and quantified by ultraviolet absorption at 275 nm (ε275 nm = 2630 $M^{-1}\,cm^{-1}$)[28].

Assay of CmoA with [U-^{13}C] chorismate. The assay mixture (0.1 ml) contained 20 mM sodium phosphate, pH 6.8, 0.2 mM SAM, 0.2 mM [U-^{13}C] chorismate and 2 µM CmoA. The reaction was incubated overnight at room temperature before analysis by mass spectrometry.

***cmoA* complementation.** For the complementation assay, the *cmoA* gene was inserted between KpnI and HindIII sites in the pQE30a (Qiagen) expression vector. *cmoA*-deficient *E. coli* cells (from the KEIO collection) were transformed with either empty vector (pQE30a), plasmid bearing wild-type *cmoA* or plasmid bearing biochemically inactive Asp89Leu *cmoA*. Transformed cells were typically grown in 50 ml lysogeny-broth media at 37 °C and induced with 0.5 mM IPTG at a $D_{600\ nm}$ of approximately 1. Cells were incubated overnight at 25 °C and collected by centrifugation. Total RNA was extracted and treated with one unit of P1 nuclease at 65 °C for 1 h. The hydrolysed nucleotide samples were analysed using LC–MS (Waters Symmetry C18 Column, 100 Å, 3.5 µm, 2.1 mm × 150 mm coupled to 12T Agilent IonSpec FT-ICR-MS) in negative mode. A linear gradient of 20% to 95% acetonitrile and 0.1% formic acid was used over 15 min. The identification of cmo5 uridine-5'-monophosphate (cmo5UMP) demonstrates the *in vivo* formation of cmo5U at the wobble position.

Cmo5UMP assay using *in vivo*- and *in vitro*-generated Cx-SAM. ^{13}C-CmoA was purified from cells grown in M9 minimal media containing ^{13}C-glucose as the sole carbon source. The media contained 1× M9 salts (Sigma), 2 mM $MgSO_4$, 0.2 mM $CaCl_2$, and 0.4% [U-^{13}C]-glucose (Cambridge Isotope Laboratory) in 0.5 l. These growth conditions yield ^{13}C-Cx-SAM bound CmoA, which was purified as described above. Notably, the occupancy of ^{13}C-Cx-SAM within ^{13}C-CmoA was nearly 100% as determined by mass-spectroscopy analysis, which is higher than that typically observed in samples prepared from cells grown in lysogeny-broth media (see Supplementary Fig. 2). ^{13}C-Cx-SAM–CmoA complex (10 µM) was added to an assay mixture containing 20 mM ammonium acetate, pH 7.3, 4 mM $MgCl_2$, 0.2 mM ^{12}C-prephenate, 0.2 mM ^{12}C-SAM, 20 µM CmoB, and total RNA isolated from CmoB-deficient cells. The assay solution was incubated at room temperature for 30 min, quenched with two units of P1 nuclease and incubated at 65 °C for 1 h. Cmo5UMP formation was analysed using LC–MS as described above (see Supplementary Fig. 12).

Network analysis. Basic local-alignment search tool (BLAST)[29] e-values for sequences in the protein family (Pfam) database[30] (see model number PF12847 (Methyltransf_18); http://pfam.sanger.ac.uk/family/Methyltransf_18) were obtained from the Structure Function Linkage Database (SFLD)[31]. SFLD BLAST searches are carried out by comparing each sequence in a superfamily to each of the other sequences. For efficiency, searches are performed by comparing bundles of 100 query sequences against all other superfamily sequences. Results are post-processed to obtain the equivalent e-value (independent of database size) based on bit scores, as would be obtained using the BLAST-2-sequences option of the BLAST+ package. Cytoscape[32] networks were created from these BLAST results at several different e-value cut-offs, and using either the full sequence set or the subset of sequences most closely related to *E. coli* CmoA (based on the BLAST e-value). The tools used to visualize protein networks were created by the University of California, San Francisco Resource for Biocomputing, Visualization, and Informatics, and are available at http://www.rbvi.ucsf.edu. Each node in the network represents a single sequence in the Pfam methyltransferase domain family and each edge represents the pairwise connection with the most statistically significant (lowest) BLAST e-value (better than the cut-off) connecting the two sequences. Connections between nodes are only shown if the e-value of the best BLAST hit between two sequences is at least as good as the specified e-value cut-off. Lengths of edges are not meaningful except that sequences in tightly clustered groups are relatively more similar to each other than sequences with few connections. The nodes were arranged using the yFiles organic layout provided with Cytoscape version 2.8. Annotation information retrieved from Swissprot[33] (functional annotation) and NCBI[34,35] (phylum), and calculated using a MUSCLE[36] multiple sequence alignment (conservation of Arg 199 in *E. coli* CmoA), was associated with each node if available.

20. Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. HKL-3000: the integration of data reduction and structure solution—from diffraction images to an initial model in minutes. *Acta Crystallogr. D* **62,** 859–866 (2006).
21. Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. Model preparation in MOLREP and examples of model improvement using X-ray data. *Acta Crystallogr. D* **64,** 33–39 (2008).

22. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
23. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
24. Dopheide, T. A., Crewther, P. & Davidson, B. E. Chorismate mutase-prephenate dehydratase from *Escherichia coli* K-12. II. Kinetic properties. *J. Biol. Chem.* **247,** 4447–4452 (1972).
25. Lorenz, M. A., Burant, C. F. & Kennedy, R. T. Reducing time and increasing sensitivity in sample preparation for adherent mammalian cell metabolomics. *Anal. Chem.* **83,** 3406–3414 (2011).
26. Kalyanaraman, C., Bernacki, K. & Jacobson, M. P. Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. *Biochemistry* **44,** 2059–2071 (2005).
27. Gibson, F. Chorismic acid: purification and some chemical and physical studies. *Biochem. J.* **90,** 256–261 (1964).
28. Parker, J. B. & Walsh, C. T. Olefin isomerization regiochemistries during tandem action of BacA and BacB on prephenate in bacilysin biosynthesis. *Biochemistry* **51,** 3241–3251 (2012).
29. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389–3402 (1997).
30. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40,** D290–D301 (2008).
31. Pegg, S. C. *et al.* Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry* **45,** 2545–2555 (2006).
32. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–432 (2012).
33. Uniprot Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40,** D71–D75 (2012).
34. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37,** D26–D31 (2009).
35. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).
36. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

TURNING POINT Engineer takes career risk in moving to biology **p.129**

NATUREJOBS BLOG The latest science-careers news and issues **go.nature.com/z8g4a7**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

IMAGES.COM/CORBIS

BIOTECHNOLOGY

Virtual reality

A growing number of biotechnology companies employ a skeleton crew of managers and outsource hands-on science.

BY HEIDI LEDFORD

If Rosana Kapeller has her way, her company will develop treatments for scourges such as cancer, cardiovascular disease and diabetes. And it will do so with only 12 full-time employees and no wet labs.

Kapeller shares a quiet office with eight colleagues at the headquarters of Nimbus Discovery in Cambridge, Massachusetts. The rest work from their homes in Missouri, Connecticut, Rhode Island and New York. This skeleton crew manages the company's operations and computer analyses; all hands-on experiments are outsourced to an international assembly of contract-research organizations (CROs). "It's a lot like managing a lab down the hall," says Kapeller, the company's chief scientific officer. "But instead of down the hall, the lab's in China and we're using Skype."

Such is life at a 'virtual' biotechnology company, a lean, nimble model that is gaining popularity among cash-hungry start-ups. These companies consist of as few as one full-time employee who oversees a drug from pre-clinical development to tests in patients, all in the hands of outside contractors.

To take advantage of this niche, scientists must have the management experience to run a remote team of researchers, and may need the financial backing to launch a company on their own. Aspirants should also be prepared for quick turnover with regard to projects and jobs: virtual start-ups are often designed to sell off individual projects — or the full company — to larger firms.

MODEL ON THE RISE

Biotechnology leaders — and their financial backers — have embraced the virtual model as a way to save money on workers and lab facilities. Nearly every biotechnology and pharmaceutical company conducts aspects of product development through contractors. But a virtual company outsources almost every step of its research and development chain.

A virtual company can be agile, shifting from drug formulation to toxicity testing without having to build facilities or hire staff. And a slimmed-down business can entice pharmaceutical companies shopping for smaller firms to restock drug pipelines.

These attributes are all the more appealing in the wake of the financial crisis, as the high risk involved in backing young biotechnology companies over the long timelines of product development makes investors wary of the ▶

sector. That pressure has already forced firms to become more efficient. "This movement is really born of necessity," says Hal Broderson, managing director of the consulting firm Rock Hill Ventures in Wynnewood, Pennsylvania. "It's like a nuclear winter out there for early-stage medical-technology companies."

But scientists interested in working for — or starting — a virtual company should also be aware of the model's limitations. Virtual companies work best when they are developing drugs for an established molecular target, using familiar techniques, cautions Kapeller. The structure is ill-suited for discovering new molecular targets, or for developing a class of drugs with a novel mode of action.

"The heart of a virtual biotech beats with the rhythm of continuous travel."
Leonide Saad

Kapeller's first company, Aileron Therapeutics in Cambridge, is developing drugs based on short helical peptides that can interact with proteins inside cells to treat diseases including cancer and endocrine disorders. Unfortunately, the approach was a little too new for the virtual model, says Kapeller, because CROs are set up to perform well-defined assays and protocols, not tackle innovative biology. Aileron survived for two years as a virtual company but eventually had to build its own wet labs and hire bench scientists. "Cutting-edge new-assay development still resides in academia, biotech and pharma," says Nancy Gillett, chief scientific officer of Charles River Laboratories, a CRO based in Wilmington, Massachusetts.

By contrast, Nimbus's structure has proved resilient thus far. The company, which has partnered with Schrödinger, a computational chemistry company based in Portland, Oregon, uses physics-driven molecular modelling to design molecules to hit cellular targets that modulate disease. CROs do the chemistry and biology studies needed to turn such molecules into viable drug candidates. Among Nimbus's 12 employees are scientists with backgrounds in biology and medicinal chemistry, who coordinate modelling efforts at Schrödinger with the hands-on work at CROs.

COMPLETE OVERVIEW

Working at a virtual biotechnology company requires a special skill set, notes David Cavalla, founder of Numedicus, a virtual pharmaceutical firm in Cambridge, UK. "You need to have somebody who has a 30,000-foot view of the whole process of drug development," he says. "They need to be able to look at the next step and say, 'This is what I'm going to need in 18 months'."

That experience is increasingly hard to come by as pharmaceutical companies and big biotechnology firms shrink their internal research and development departments, laying off scientists and outsourcing their efforts. Increasingly, drug-development jobs are to be found at CROs rather than at classical, integrated biotechnology firms. Gillett says that when she left Genentech, a large biotechnology company based in South San Francisco, California, to join a small CRO, people told her she was committing career suicide.

That was nearly 20 years ago, when CROs were seen as employers of last resort for scientists, paying less and offering less autonomy than jobs at pharmaceutical companies. Since then, things have changed dramatically, says Gillett. "Now the big companies are coming to us for advice."

Scientists at a CRO may gain experience from working on many different projects, and can advise clients on specific areas of drug development. But they rarely get to participate in strategic decision-making about the direction of a project, or develop the overarching view of the process that Cavalla advocates. Senior scientists who have left big pharma, or have been laid off, remain a key source of management experience, he says. "The reason you're able to make this virtual model work is because you can hire all of these experienced grey-hairs from pharma companies."

David Collier, managing director for life sciences at CMEA Capital in San Francisco, argues that some young scientists will still be able to find training at the remaining big firms. While there, he notes, they can seek out the experience most needed in a virtual company: managing outside contractors. "The key part is to understand how a CRO works and how to negotiate a reasonable price," he says. Such a level of experience includes everything from designing contracts to ensure that contractors stick with the company timeline, to making sure that basic lab protocols are up to standard.

Ilyas Washington invented a treatment for blindness that launched a virtual company.

MICHAEL GHARBI

That does not mean that being in charge is all management and no science. The people best positioned for success in a virtual biotech combine management experience with scientific acumen, says Leonide Saad, founder, president and sole full-time employee of Alkeus Pharmaceuticals in Boston, Massachusetts. Saad, a tissue engineer by training, says that running a virtual company frees him from internal bureaucracy so that he can spend more time considering the bigger scientific picture. "It's a blast when you're on your own," he says. "When everything is in-house, you spend much more time managing people rather than thinking about your core drug and your core development."

The virtual model made it possible for Saad to strike out on his own by reducing the cost of launching a company, but investors will still want firm evidence of success before they will risk their cash. Saad had been a venture capitalist for two years when he decided to launch his own company with start-up funding borrowed from family members and his own savings.

His plan was to seek further investment once he had something to show. "If you're an entrepreneur you need to have cash to survive and build value for a full year," he says. "You cannot survive on ramen noodles and then go to venture capitalists and say 'I haven't made progress because I don't have your money'."

THE RIGHT CHOICE

Saad knew that he needed a project that could prove its worth on a limited budget in about two years — before his money ran out. He trawled through more than 150 university patents in search of a technology that he could build his company around, evaluating each with the eye of a venture capitalist. For a situation like his, he says, it was important to seek out projects that were focused, with a clear and preferably short path to therapeutic application.

Saad narrowed the list down to 20 technologies, then researched the intellectual property to determine whether the patents were strong enough to hold up if challenged in court. He also read up on the literature to see whether a scientific consensus was building in support of the proposed invention. Finally, he settled on a possible therapy for macular degeneration — a common cause of blindness — created by Ilyas Washington, an ophthalmology researcher at Columbia University in New York. Saad now spends his days visiting Washington and the five CROs that are working on the project. "The heart of a virtual biotech beats with the rhythm of continuous travel," he says. "I carry the entire office with me on a laptop."

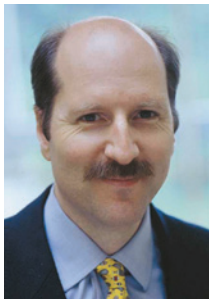
Not everyone is so enamoured with the virtual lifestyle. Stewart Lyman, owner of Lyman BioPharma Consulting in Seattle, Washington, worries that the trend leaves few satisfying research jobs in drug discovery.

Although CROs are booming, he says, jobs

with them will not fulfil many of the best researchers, who prefer to have scientific control over their work rather than doing the bidding of a client. Scientists seeking a career in CROs should look for jobs that will give them autonomy, agrees Jonathan Montagu, vice-president for business at Nimbus. "You have to be selective."

DIVIDE AND CONQUER

Scientists who find jobs at virtual companies should recognize that they may soon be back on the job market. Virtual firms are often designed to be bought by pharmaceutical companies, giving investors a chance to recoup their funds without waiting for the decade or more that it can take to bring a drug to market. "If I were a young scientist, a virtual biotech is not the kind of place I would aspire to work," says Lyman. "Even if you're successful, you're going to get liquidated in a couple of years and then you're out of a job again."



"You need to have somebody who has a 30,000-foot view of the whole process of drug development."
David Cavalla

Nevertheless, job seekers looking for stability may find options in a new breed of virtual biotechnology company. Some companies are structured to enable the sale of individual projects, leaving the rest of the firm intact and allowing employees and infrastructure to remain in place. Nimbus, for example, has set up each of its projects as a separate subsidiary with intellectual property and assets that a pharmaceutical company can acquire without buying the full firm.

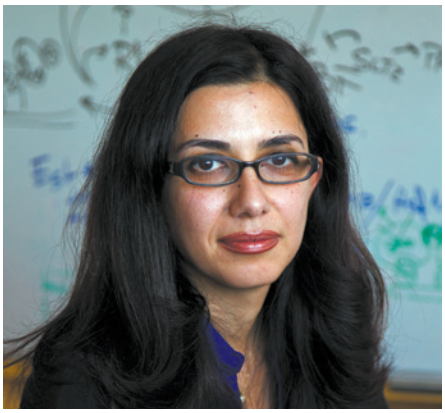
Similarly, when Collier and his colleagues launched Velocity Pharmaceutical Development, based in La Jolla, California, they configured each project as its own corporation. "There is a lot of experimentation now with new models," says Collier.

Ultimately, industry scientists need to adapt to the new normal, says Justin Chakma, an analyst at venture-capital firm Thomas, McNerney & Partners in La Jolla, California. That may mean jumping from job to job. "Scientists need to be comfortable working almost as consultants," he says. "It's not a steady stream of income like it was years ago." ■

Heidi Ledford *reports for* Nature *from Cambridge, Massachusetts.*

TURNING POINT
Hana El-Samad

Trained as an engineer, Hana El-Samad honed her skills to study complex systems, but ended up researching gene expression in real time. This year, she won a US$1.4-million grant from the Paul G. Allen Family Foundation in Seattle, Washington. Unbound by conventional biology instruction, El-Samad feels free to take risks.



Who influenced you to pursue science?
I grew up in Lebanon, where my mother, a maths teacher, instilled in me a love of maths and engineering. At the American University of Beirut, I wanted to study mathematical theories of how things work. I focused on control theory, which looks at automated systems.

How did you shift into biology?
In 1999, I earned a master's degree in electrical engineering with a focus on controlled dynamic systems from Iowa State University in Ames. In 2002, halfway through my PhD, my adviser, Mustafa Khammash, moved to the University of California, Santa Barbara, and I went with him. People were starting to talk about systems biology, and I realized that the theories I had been studying on machines would be relevant to systems created by nature. I completely switched gears.

Did your adviser support that?
He is smart and open-minded, and thought that tackling biology could be really interesting. We divided up chapters of a biology book and taught each other. My thesis was on heat-shock responses that bacteria use to adapt to temperature increases. We tried to model them to understand how they operate in real time.

Was it a difficult to move into biology?
In 2004, I earned my PhD in mechanical engineering. I faced a choice — accept an engineering position or throw myself into biology. It was not a trivial decision. I had offers for several engineering posts, but a collaborator had nominated me for the Sandler Fellows Program at the University of California, San Francisco (UCSF), which funds one person each year to start a small, independent group focused on risky research. I chose that. People thought I was crazy, but it was the best thing for my career. I'm now an experimental scientist, a hybrid of an engineer and a biologist.

How did you expand your lab?
I didn't want a gigantic lab: I wanted six to eight people who do thorough, in-depth science, to try to understand how a small number of systems work in predictive ways. I chose people with backgrounds in maths, physics, molecular biology and computational science.

Did that set-up have challenges?
Yes — piecing together a mosaic of disciplines left us with no common language. At early lab meetings, I wanted to pull my hair out. People were talking about the same thing using different terminology, and getting frustrated. There was also a reluctance to ask what might be considered stupid questions.

How did you get past those barriers?
I wrote a lab constitution that acknowledges that we are all from different backgrounds, that we shouldn't all be expected to understand everything — and that we should ask questions. It is written playfully and we update it as necessary.

Is it hard to get federal grants for your multidisciplinary research?
It can be. I believe that agencies want to fund this kind of science, but they have to funnel grants through review groups that can have conservative reviewers. Still, we did get a US National Institutes of Health grant in 2010 to fund the UCSF Center for Systems and Synthetic Biology.

You have two grants from private foundations. Why do you think your work appeals to them?
The David and Lucile Packard Foundation in Los Altos, California, liked our approach to cell-to-cell variability, and the Allen Foundation liked how we decided to decipher the genetic encoding and decoding that allow cells to survive in complex environments. I think both like to fund high-risk, potentially transformative things that are not necessarily attractive to agencies, and we don't do run-of-the-mill stuff. ■

INTERVIEW BY VIRGINIA GEWIN

RONDO CODE

In perfect harmony.

BY TONY BALLANTYNE

"I remember, I was teaching kids how to program computers. I was trying to think of ways to make it easier for them. The thing is, they have no trouble writing lists of instructions; what they find difficult to understand is the looping and the branching…"

Ada broke off as the sound of an orchestra sprang up and the whole world paused. The traffic in the road by the little pavement café, the pedestrians, even the birds in the lime trees. Somewhere up in the sky, aeroplanes hung motionless for a moment. And then their courses adjusted slightly, the music came to an end and the world resumed. Everyone relaxed. Four days since the big glitch, and everything seemed back to normal.

"You understand what I mean?" she said. The journalist opened his mouth to answer and she interrupted him. "Of course you do *now*, but before all the changes, this is what kids used to struggle with."

Ada had taken a dislike to the journalist. He had arrived at the interview with his mind already made up. She was amusing herself by not giving him a chance to speak.

"Like you can draw a square by repeating four times the routine *go forward ten steps and turn right ninety degrees*. That's an example of a loop. Kids used to struggle with that. Adults used to struggle with that.

"So I was trying to think of a way to help people *understand*. I was listening to kids singing *The Twelve Days of Christmas*, and I thought *that's really quite complex*: in the old jargon, the song is an example of nested loops. You count from one to twelve, first day of Christmas, second day, yes? And then for each day you have to count backwards to one: seven swans a-swimming, six geese a-laying, five gold rings. I thought, that's not a song, that's a code structure.

"I thought, they understand songs, maybe I could teach them to code that way. That's when I came up with the idea of Rondo Code."

She sipped at her drink.

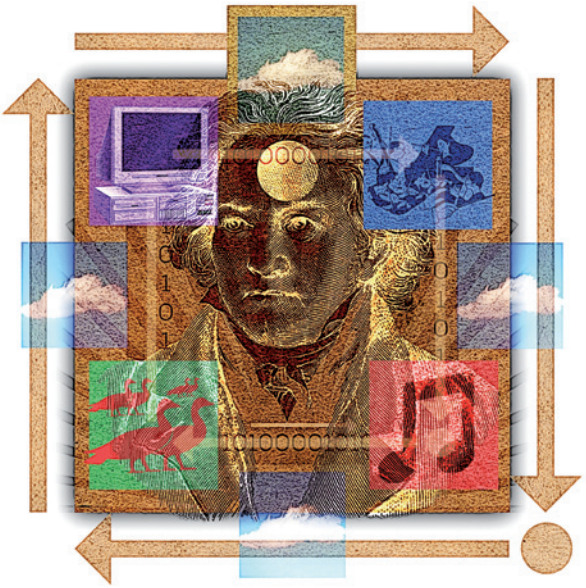
"Why's it called that?" asked the journalist, free to speak at last.

"Rondo Code? I thought everyone knew that by now, Mr Leibniz. A rondo is a musical form. You play tune A, then tune B, then tune A, then tune C, then tune A and so on. It was a catchy name that sort of described what I was attempting."

Mr Leibniz smiled and Ada knew that the piece of information was of no interest. He wanted to assign blame. People had nearly died in the big glitch.

"Well, it worked," he said. "Anyone who can sing a tune can now understand how to program a computer."

"It's not just that. Think of all the testing we used to do, all of that debugging. Now you can tell if a program is well written just by listening to it. Does it have musicality?"



JACEY

"That's an interesting philosophical point," said the journalist. "Are humans programmed to program?"

"It's a silly point," said Ada. "Anthropomorphic thinking. We do what we do."

"Hmm," said the journalist. "But of course, all that was a prelude. Your stroke of genius was still to come."

"Nonsense. Nothing I have done could be described as genius. Rondo Code was a good idea, that is all. There was a lot of hard work went into the syntax and structure. I have shown dedication, nothing more."

"Others might disagree, Ms Byron. Cottrel says your idea to put existing music through Rondo Code was genius."

"Cottrel is a second-rate composer, not a programmer."

"Where did you get the inspiration from, Ms Byron?"

"I hate that word, inspiration. If I hadn't thought of the idea, someone else would have."

"But they didn't, Ms Byron." He shook his head in wonder. "Who would have thought it? That Beethoven's Sixth Symphony could predict the weather? Or that Wagner's *Tristan und Isolde* could control negotiations between warring states?"

"That program has yet to bring a satisfactory resolution to a conflict," said Ada, tersely.

"But it keeps both parties in dialogue rather than fighting. You've got Duke Ellington running primary education and Melissa Hui controlling the traffic…"

"I think you're simplifying things for your readers, Mr Leibniz. The music needed some adaptation…"

Leibniz waved a hand, and it was obvious to Ada that those words would never see print either.

"Was it your idea to use Mozart to control well-being?"

Ada was silent for a moment. Here was the blame.

"No. Nor was I the hero who managed to correct the code and allow us all to sleep again."

Leibniz stared at her. He wasn't quite ready to give up.

"Why did you agree to this interview, Ms Byron? You're famous for being something of a recluse."

"I just don't like talking to the press. It may surprise you to learn that's not the same thing."

The journalist laughed.

"So why speak now?"

"I just wanted people to understand. The big glitch is over."

"Hmm." He tapped his pencil on his teeth. "So, my final question. You're not universally popular, are you? It has been said that once a piece of music is put to work as Rondo Code, all the pleasure is taken from it."

"I've heard that said," said Ada. "There are fools in every age."

"But surely, once a piece of music has been reduced to a mechanical series of notes, once it has been fully understood by a machine, surely the pleasure has all evaporated."

"Since when did understanding spoil pleasure?" asked Ada, standing up. "In my experience, it tends to enhance it." ■

Tony Ballantyne *has written many short stories. His sixth novel,* Dream London, *will be published in October by Solaris.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

Intuition and cooperation reconsidered

ARISING FROM D. G. Rand, J. D. Greene & M. A. Nowak *Nature* **489,** 427–430 (2012)

Rand *et al.*[1] reported increased cooperation in social dilemmas after forcing individuals to decide quickly[1]. Time pressure was used to induce intuitive decisions, and they concluded that intuition promotes cooperation. We test the robustness of this finding in a series of five experiments involving about 2,500 subjects in three countries. None of the experiments confirms the Rand *et al.*[1] finding, indicating that their result was an artefact of excluding the about 50% of subjects who failed to respond on time.

There are two major problems in the analysis by Rand *et al.*[1] First, their exclusion of subjects who fail to respond on time cause a selection problem. In their observational studies, Rand *et al.*[1] show that slow responders cooperate less. The exclusion of slow responders therefore automatically increases cooperation in the time-pressure treatment. Second, when including all subjects in the Supplementary Information analyses, they incorrectly control for whether subjects answer on time. Without controlling for this endogenous variable, the time pressure effect is not significant (t value = 1.62, P value = 0.11 in both study 6 and study 7).

We test the robustness of the Rand *et al.*[1] results in experiments 1–4, and experiment 5 is a replication. To minimize missing values we use a binary decision. In experiment 1, two subjects simultaneously decide whether to keep X or give a larger amount to the other individual in a prisoner's dilemma[2]. X is varied in five rounds with new pairs in each round. Subjects (Swedish students, $n = 167$) are randomly allocated to deciding within 10 s or waiting 10 s before deciding. The mean rate of cooperation is about 50% in both groups (t value = 0.33, P value = 0.740) (Fig. 1a).

The maximum time to respond is reduced to 7 s in experiments 2–4, and a four-person public goods game is used[2]. Subjects decide whether to keep a fixed amount or give a larger amount to the group (the amount is varied in four rounds with new groups in each round). Experiment 2 (Swedish students, $n = 199$) and experiment 3 (USA general population sample, $n = 583$) have identical designs. In experiment 4 (Austrian students, $n = 320$), the time subjects have to wait before deciding is increased to 20 s and the wording is changed slightly. The time pressure effect is in the opposite direction of Rand *et al.*[1] in experiments 2–4, but not significant (t value = −0.55, P value = 0.586 in experiment 2; t value = −0.44, P value = 0.663 in experiment 3; t value = −1.93, P value = 0.054 in experiment 4) (Fig. 1a). Pooling experiments 1–4, the rate of cooperation is 44% with time pressure and 47% without time pressure (t value = −1.29, P value = 0.197). Including only the first round decision, the rate of cooperation is 44% with time pressure and 46% without time pressure (Chi-square = 0.60, P value = 0.432).

In experiments 1–4, subjects knew that they would be making decisions under time pressure before they reached the decision screen. In Rand *et al.*[1], subjects did not know about the time pressure until they reached the decision screen. Rand *et al.*[1] also included an example in the instructions, but the example may also prime decisions (the example ended with "Thus you personally lose money on contributing"). These differences are tested in experiment 5 in a one-shot public goods game with six treatments. Treatments 1 and 2 are a replication of the Rand *et al.*[1] design, but with a binary decision. Treatments 3 and 4 are identical to treatments 1 and 2, but do not include the example. Treatments 5 and 6 replicate treatments 3 and 4, but information about time pressure is given before the decision screen. Design and wording of experiment 5 were done in collaboration with D. Rand. Data are collected on Austrian students ($n = 353$), and two USA general population samples (Decision Research sample, $n = 251$; Qualtrics Panels sample, $n = 600$).

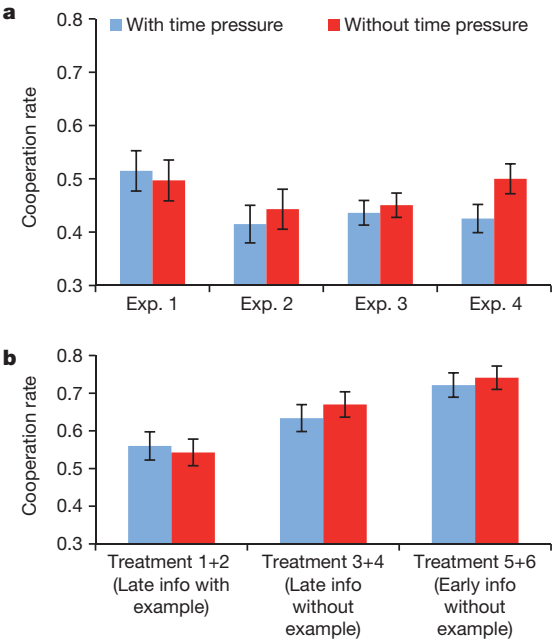


Figure 1 | Time pressure does not increase cooperation in social dilemmas. **a**, Mean (± s.e.) rate of cooperation with and without time pressure in a repeated prisoner's dilemma game with stranger matching (experiment 1, Exp. 1) and a repeated public goods game with stranger matching (experiments 2–4). The result in Rand *et al.*[1] of time pressure increasing cooperation is not confirmed in any of the experiments. **b**, Mean (± s.e.) rate of cooperation in treatments 1–6 in a one-shot public goods game (experiment 5). The result in Rand *et al.*[1] of time pressure increasing cooperation is not confirmed in any of the comparisons. Treatments 1 and 2 replicate the Rand *et al.*[1] design (with late information that the decision will be made under time pressure), but with a binary decision. Treatments 3 and 4 are the same as treatments 1 and 2, but without the example used by Rand *et al.*[1] Treatments 5 and 6 replicate treatments 3 and 4, but provide early information that the decision will be made under time pressure (as in experiments 1–4). The rate of contribution is lowest in treatments 1 and 2 consistent with a priming effect of the example.

No significant effect of time pressure for any of the comparisons is found (Fig. 1b). The rate of cooperation is 56% with time pressure and 54% without time pressure in the replication of Rand *et al.*[1] (Chi-square = 0.11, P value = 0.737). The most striking result is that including the example reduces cooperation, consistent with a priming effect (Chi-square = 8.16, P value = 0.004).

We conclude that forcing individuals to decide quickly in social dilemmas does not in general increase the rate of cooperation, casting doubt on the Rand *et al.*[1] interpretation of humans as intuitively cooperative.

METHODS

In the five rounds of experiment 1, the subjects decide between giving SEK150 to the other player and keeping between SEK40 and SEK90. In the four rounds of experiment 2/3/4, the subjects decide between keeping SEK50/€5/$2.5 and giving between SEK75–150/€7.5–15/$3.75–7.5 to the group. In experiment 5, the subjects decide between keeping an amount ($2 in the Decision Research sample, $4 in the Qualtrics Panels sample, and €4 in the Austrian sample) and giving twice

as much to the group. An Appendix with more detailed descriptions of the methods and results are available from the authors.

Gustav Tinghög[1,2], David Andersson[1], Caroline Bonn[3], Harald Böttiger[4], Camilla Josephson[1], Gustaf Lundgren[5], Daniel Västfjäll[6,7], Michael Kirchler[3,8] & Magnus Johannesson[1,5]

[1]Division of Economics, Department for Management and Engineering, Linköping University, SE-581 83 Linköping, Sweden.
[2]The National Center for Priority Setting in Health Care, Department of Medical and Health Sciences, Linköping University, SE-581 83 Linköping, Sweden.
[3]Department of Banking and Finance, University of Innsbruck, Universitätsstrasse 15, 6020 Innsbruck, Austria.
[4]Klarna AB, Norra Stationsgatan 61, SE-113 43 Stockholm, Sweden.
[5]Department of Economics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden.
e-mail: magnus.johannesson@hhs.se

[6]Department of Behavioural Sciences and Learning, Linköping University, SE-581 83 Linköping, Sweden.
[7]Decision Research, 1201 Oak Street, Suite 200, Eugene, Oregon 97401, USA.
[8]Centre for Finance, Department of Economics, University of Gothenburg, Box 600, SE-40530 Göteborg, Sweden.

Received 6 February; accepted 16 April 2013.

1. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489,** 427–430 (2012).
2. Camerer, C. F. *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, 2003).

Author Contributions G.T. and M.J. designed research; all authors performed research; D.A. analysed data; G.T., M.K. and M.J. wrote the paper.

Competing Financial Interests: Declared none.

doi:10.1038/nature12194

Rand *et al.* reply

REPLYING TO G. Tinghög *et al. Nature* **498,** http://dx.doi.org/10.1038/nature12194 (2013)

Tinghög *et al.*[1] take issue with two of the ten experiments in our paper[2] (studies 6 and 7). Here we reanalyse the data from these experiments as suggested by Tinghög *et al.*[1], and demonstrate that our reported positive effect of time pressure on cooperation is not an artefact. Furthermore, an aggregate analysis based on fifteen studies and 6,910 decisions also replicates this effect[3].

In studies 6 and 7, we examined the relationship between intuition and cooperation by manipulating decision time: in one condition, subjects playing a public goods game were asked to decide in less than 10 s; in the other condition subjects were asked to think for at least 10 s before deciding. Tinghög *et al.*[1] make an excellent point regarding potential issues related to excluding subjects who did not respond in time (or including a dummy variable controlling for failure to obey the time constraints).

Here we reanalyse our data following the suggestions of Tinghög *et al.*[1]. We do not exclude subjects who failed to obey the time constraint, and we do not control for such failure. As in our original analyses, we find a significant positive effect of time pressure on cooperation ($N = 891$; rank-sum, $P = 0.014$; Tobit regression with demographic controls, $P = 0.022$; we combine studies 6 and 7 because of a non-significant interaction between time pressure and study, $P = 0.62$). Thus the time-pressure effect reported in our previous paper is not an artefact of exclusion, as Tinghög *et al.*[1] have suggested.

Furthermore, our original paper presented data from ten studies using three distinct methods to test whether people's automatic, intuitive responses are more or less cooperative than responses generated through reflection and deliberation. All of these studies supported the conclusion that, on average, intuition favours cooperation. The concerns of Tinghög *et al.*[1] apply to two of these ten experiments, which used only one of our three methods. They do not challenge the convergent evidence presented by the other eight studies. On the contrary, their criticism of our studies 6 and 7 is based on their acceptance of the correlational results we reported in studies 1–5.

Tinghög *et al.*[1] report five experiments in which there is no significant effect of time pressure on cooperation. However, four of these experiments involve design changes that are likely to eliminate the time-pressure effect. First, in these experiments subjects played the cooperation games after having made a series of other economic decisions. Thus they had been given an opportunity to adjust to the laboratory setting, reducing the spillover of intuitions from outside the lab. As demonstrated in our study 9, previous experience eliminates the positive effect of intuition[2]. Second, subjects were under time pressure not only when deciding, but also when acquiring information about the payoff structure. As noted in our Supplementary Information[2], faster acquisition of payoff information is associated with decreased cooperation[4]. This is because cooperative decisions require information about the payoffs to others, rather than just one's own payoff. Thus, the confounding of these opposite effects of time pressure on information acquisition and prosociality would be expected to result in a null effect.

In the fifth experiment of Tinghög *et al.*[1], these problems are eliminated. This study's null result is disappointing. However, it fits within the pattern of results observed in an aggregate analysis examining every experiment our group has ever run applying time pressure to social dilemmas (thus eliminating potential "file drawer" effects)[3].

Across 15 studies and 6,910 decisions, there is a highly significant positive effect of time pressure on cooperation. This effect persists when including subjects who did not obey the time constraint. Furthermore, there is substantial study-to-study variation, with some studies showing significant positive effects of time pressure and others showing no effect. Critically, no study shows a significant negative effect of time pressure on cooperation, consistent with the null (but non-negative) results of Tinghög *et al.*[1]. We also find that, over the last 2 years, the size of the time-pressure cooperation effect has steadily decreased in the subset of studies run on Amazon Mechanical Turk (AMT[5]). Given the marked increase in the popularity of AMT as a platform for behavioural experiments, this is consistent with our previous finding that experience undermines the intuitive cooperation effect.

In sum, our findings are supported by (1) a reanalysis of studies 6 and 7, (2) the remaining eight studies reported in our original paper, and (3) an aggregate analysis of data from over a dozen other time-manipulation studies. Thus, there is clear convergent evidence that intuition promotes cooperation on average, but not in all cases, nor

for all people. Deepening our understanding of the factors that moderate the effect of intuition on cooperation is an important direction for future research, one that we hope Tinghög and collaborators will join us in pursuing.

David G. Rand[1,2,3], Joshua D. Greene[2] & Martin A. Nowak[1,4,5]
[1]Program for Evolutionary Dynamics, Harvard University, Cambridge, Massachusetts 02138, USA.
e-mail: drand@fas.harvard.edu
[2]Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, USA.
[3]Department of Psychology, Yale University, New Haven, Connecticut 06520, USA.
[4]Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138, USA.
[5]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

1. Tinghög, G. *et al.* Intuition and cooperation reconsidered. *Nature* **498,** http://dx.doi.org/10.1038/nature12194 (2013).
2. Rand, D. G., Greene, J. D. & Nowak, M. A. Spontaneous giving and calculated greed. *Nature* **489,** 427–430 (2012).
3. Rand, D. G. *et al.* Intuitive cooperation and the social heuristics hypothesis: evidence from 15 time constraint studies. Preprint at SSRN http://ssrn.com/abstract=2222683 (2013).
4. Piovesan, M. & Wengström, E. Fast or fair? A study of response times. *Econ. Lett.* **105,** 193–196 (2009).
5. Rand, D. G. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* **299,** 172–179 (2012).

Author Contributions D.G.R., J.D.G. and M.A.N. performed the analysis and wrote the paper.

doi:10.1038/nature12195